UNIVERSIDAD MICHOACANA DE SAN NICOLÁS DE HIDALGO

FACULTAD DE QUÍMICO FARMACOBIOLOGÍA

"Estudio QSAR por Algoritmos Genéticos de Reconocedores de Surco del DNA"

TESIS

Que presenta

Anai Zavala Franco

Para obtener el título profesional de

QUÍMICO FARMACOBIÓLOGO

Asesor de Tesis

D.C. Luis Chacón García

Morelia, Michoacán Octubre de 2012

El presente trabajo se realizó en el laboratorio de Diseño Molecular del Instituto de Investigación Químico – Biológicas de la Universidad Michoacana de San Nicolás de Hidalgo con la asesoría del D.C. Luis Chacón García.

Se agradece el apoyo de la Coordinación de Investigación Científica de la UMSNH (Proyecto 2.18).

Una parte de este trabajo se presentó en la 7ª Reunión de la Academia Mexicana de Química Orgánica realizada del día 4 al 8 de abril de 2011, en la ciudad de Cuernavaca, Morelos.

La primera parte de esta investigación se publicó en la Revista Biológicas, de la DES Ciencias Biológicas Agropecuarias de la UMSNH, en el volumen 12, número 2, páginas 108 – 115, en diciembre de 2010 bajo el nombre de “Estudio QSAR por algoritmos genéticos de reconocedores de surco del DNA” (Se anexa copia al final del trabajo).

Dedicado a Juana, mi madre y a Jaime, mi padre.

Por todo lo que han hecho por mí, simplemente por darme vida.

A mis hermanas, que siempre están ahí.

Gracias Dios, por ponerme en este mundo, en este momento,

en este lugar y con estas personas. Gracias por darme

la fuerza necesaria cuando me quedo sin ella.

AGRADECIMIENTOS

Gracias es lo único que se puede dar a aquellas personas que te rodean y que de forma directa o indirecta, contribuyen para el crecimiento personal, profesional y espiritual de uno mismo.

Gracias, Doctor Luis Chacón, por proponerme el proyecto y confiar en mí para realizarlo, por darme una oportunidad única y contribuir a mi desarrollo profesional, además de la amistad que me ha brindado.

Gracias, D.C. Betzabe, D.C. Surizaddai, D.C. Claudia, M.C. Sandra y M.C. Yolanda, por todas las aportaciones que tuvieron a bien proponer para el enriquecimiento y presentación de este trabajo.

Gracias BM's, Jessica, Aracely, Wendy, Mario, César y Chava, sin ustedes la escuela (y ya la vida), no sería la misma. También a Yared, Selene, Ale y Marlene, que en poquito tiempo hicimos una buena y bonita amistad.

Gracias, Nancy y Ana Laura, por que siempre están ahí aunque yo me vaya, me regañan, me alientan y me apoyan. No se mueran nunca, las quiero.

Gracias, Natalia, siempre estás ahí para todo, sin importarte nada más. Siempre formarás parte de esto y de lo que viene. Peri, gracias por llegar a mi vida, gracias por todo lo que me has dado, tu amistad es muy valiosa.

Gracias Luis, por ser tú y estar ahí cuando lo necesité. Gracias por todo.

Gracias a todos mis amigos, por que han contribuido de una u otra forma a lo que soy ahora, los quiero mucho: Luis, Karla, Magaly, Eddie y Rafa.

Gracias a mis compañeros de laboratorio por hacerlo tan agradable: Ana Lilia, gracias por enseñarme la técnica. Juanita, ya llegó la hora y estás ahí. Rosy, siempre ayudándome a todo. Ale, la confianza depositada en mí hace que me esfuerce mucho más.

Gracias Araceli, Liz, Alba e Inés, por brindarme su amistad y apoyo.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ABREVIATURAS

QSAR	Estudio de relación cuantitativa estructura-actividad.
DNA	cido desoxirribonucleico.
D_{CL}	Descriptor topológico.
Ec	Energía de acoplamiento.
AM1	Método semiempírico Austin Model 1.
PM3	Método semienpírico Parametrization Method 3
MNDO	Método semiempírico Modified Neglected Diatomic Overlap
A	Adenina
T	Timina
G	Guanina
C	Citosina
DDD	Dodecámero de Dickerson-Drew
$E^{\text{DNA-LIG}}$	Energía de complejo DNA- ligando.
E^{CPLX}	Energía potencial absoluta del complejo final
E^{DBC}	Energía de un solo punto en la conformación que adopta el DNA en el complejo final.
E^{LFC}	Energía del ligando en la conformación final.
$\Delta E^{\text{LIGANDO}}$	Energía de distorsión del cambio de conformación del ligado del estado mínimo libre de la conformación del farmacóforo.
ΔE^{DNA}	Energía de distorsión del cambio de conformación del DNA del estado mínimo libre de la conformación del farmacóforo.
log (ΔTm)	Logaritmo de la variación de la temperatura de desnaturalización del DNA
Tm	Temperatura de desnaturalización del DNA.
nm	nanómetros

AG	Algoritmo genético
Q^2	Capacidad predictiva del modelo.
R^2	Indicador de la precisión del modelo (Coeficiente de determinación).
Q^2boot	Coeficiente de determinación de validez cruzada.
F	Parámetro de Fisher.
s	Desviación estándar.
Kx	Correlación.
n	Número de datos
°C	Grados Celsius
cos	Coseno
kcal	Kilocalorías
mol	Moles
Å	Amstrong
DAPI	4',6-diamidino-2-fenilindol
PDB	Protein Data Bank
AP	Dipéptido alanil-fenilalanina (1-4)

1. INTRODUCCIÓN

La inteligencia artificial se ha convertido en una herramienta muy útil en la tecnología moderna y en parte de las ciencias como la síntesis química y predicción de actividad biológica, las cuales han sido empleadas en el descubrimiento de nuevos fármacos para tratar las enfermedades crónicas graves que se presentan hoy en día, dicha inteligencia artificial utiliza una serie de ecuaciones matemáticas donde se simula el ambiente propio de actividad de cada molécula a estudiar, midiendo todas las características posibles y así cuantificar parámetros útiles. Estas características adoptan el nombre de descriptores moleculares.

En este trabajo se describe un estudio de relación cuantitativa estructura-actividad QSAR (Quantitative Structure Activity Relationship) por algoritmos genéticos de una serie de 27 bisamidinas aromáticas, de las cuales se conoce su capacidad como reconocedores de surco menor del DNA, esta característica les permite poseer actividad antiparasitaria, antibacteriana y citotóxica. Los modelos obtenidos son estadísticamente significativos y contienen al descriptor D_{CL}, un descriptor de características topológicas que mediante su aplicación junto a descriptores geométricos y de electronegatividad, predicen teóricamente la actividad de un péptido propuesto como posible reconocedor de surco del DNA.

También, mediante un estudio Docking, se describe un acoplamiento del péptido propuesto con el DNA y se obtiene la energía de acomplejamiento (E_c) óptima. Con ello, se da mayor solidez a la utilidad del descriptor D_{CL} como una característica importante en el reconocimiento de compuestos que interaccionan con el DNA.

2. ANTECEDENTES

Durante la década de los setenta apareció un campo nuevo del conocimiento donde se daba un uso a la computadora, de manera relativamente accesible, de tal forma que se enfocó al estudio teórico de moléculas y también al diseño de estructuras novedosas con posible actividad biológica. A esta nueva disciplina se le conocería como Química Computacional. Las grandes compañías farmacéuticas fueron las mayores impulsoras del campo y se dio un giro total a la química; ya no era la ciencia experimental que se creía, pues no era necesario hacer experimentos para obtener un resultado. Actualmente, la predicción teórica de propiedades químicas rivaliza, incluso con ventaja, con determinaciones experimentales. La química computacional se basa en el uso de modelos matemáticos para la predicción de propiedades químicas y físicas de compuestos, utilizando computadoras. Se determina la estructura y propiedades moleculares mediante el uso de mecánica molecular, métodos semiempíricos y/o teorías de orbitales moleculares.

El objetivo de la química computacional es modelar la química experimental, pero cada nivel de aproximación tiene sus limitaciones inherentes. Por fortuna, los errores sistemáticos se pueden eliminar cuando se comparan con las propiedades de las moléculas y conforme los modelos se vuelven más sofisticados se aproximan más a la química práctica.

El uso de la química teórica modelo se basa en el principio que dicta: "*Un modelo teórico debe ser aplicable en forma uniforme a cualquier sistema molecular, independientemente de su tipo y tamaño, siendo la capacidad de cómputo la única*

limitante que debe imperar"[1]. La implementación de los modelos teóricos al cómputo es denominada química teórica modelo, o simplemente modelo químico. Cuando un modelo es capaz de reproducir los resultados conocidos, entonces se le puede emplear para predecir propiedades de otros sistemas.

2.1 Técnicas de simulación molecular

Los métodos que permiten obtener información a nivel molecular del sistema en estudio se agrupan en dos métodos: los métodos cuánticos y los métodos clásicos. Los primeros aluden al sistema molecular mediante un conjunto de núcleos y electrones que siguen las leyes fundamentales de la Mecánica Cuántica. Los métodos clásicos describen el sistema como un conjunto de partículas elementales, localizadas sobre los núcleos y cuyas interacciones se aproximan a una suma de términos energéticos representados por expresiones basadas en la Mecánica Clásica. El método clásico describe de forma menos rigurosa el sistema, pero reduce los gastos computacionales.

Las técnicas cuánticas ofrecen más confianza y rigurosidad, permiten obtener de forma fiable propiedades como la geometría o la energía de sistemas químicos y ya que consideran los electrones explícitamente, pueden cuantificar las propiedades relacionadas con esto (parámetros espectroscópicos, distribución de cargas, momento dipolar, etc.), pero las exigencias computacionales son muy altas y sólo pueden ser aplicadas a sistemas pequeños (pocos átomos). La Mecánica Cuántica considera una función de onda que contiene toda la información posible

del sistema. Para obtener esta función es necesario resolver la ecuación de Schrödinger:

$$H\psi = E\psi \qquad \textbf{Ec. 1}$$

Donde H incluye la energía cinética y potencial de núcleos y electrones, y E la energía del sistema.

Para poder estudiar sistemas más grandes se introducen simplificaciones, como la representación de la función de onda en términos de orbitales moleculares y la expresión de estos como combinación lineal de orbitales atómicos. Existen tres metodologías básicas para la obtención de la función de onda: *ab initio*, la semiempírica y la basada en la Teoría del Funcional de la Densidad.[2]

Métodos *ab initio*. No es necesario determinar parámetros empíricos, es teoría solamente. Tales métodos sin embargo, son muy lentos y resultan prohibitivos para sistemas de tamaño medio.

Métodos semiempíricos: Emplean determinaciones de parámetros semiempíricos y son más atractivos que la teoría pura. A diferencia de los métodos *ab initio*, la exactitud de algunos métodos semiempíricos se limita a la exactitud de los datos experimentales usados en la obtención de los parámetros. Sin embargo, los métodos semiempíricos son suficientemente rápidos y precisos para su aplicación rutinaria en sistemas bastante grandes.[3] Entre los métodos semiempíricos más populares cabe destacar el Austin Model 1 (AM1), el Parametrization Method 3 (PM3) y el Modified Neglected Diatomic Overlap (MNDO).

2.2 Método semiempírico AM1

Modelo Austin 1, nombrado así en honor a la Universidad de Texas en Austin, creado por el grupo de Dewar[4]. Ellos intentaron crear un método que fuera aplicable a sustancias de interés biológico, para lo cuál debería predecir la existencia de puentes de hidrógeno.[5]

2.3 Métodos computacionales de modelado molecular y diseño de fármacos.

Cualquier acción farmacológica tiene su inicio en la formación de un complejo entre la molécula de fármaco y su sitio receptor en una macromolécula biológica. Así, la especificidad de la respuesta a un fármaco se determina por la capacidad de los receptores celulares para reconocerlo y provocar o no una respuesta. Como consecuencia de esa unión, se pueden hacer posibles manipulaciones con fines terapéuticos. El receptor debe ser contemplado como un sitio selectivo, capaz de distinguir entre posibles ligandos y moléculas que no poseen afinidad de unión. Además, es el primer responsable de la serie de acontecimientos que pueden traducir esa interacción en una respuesta celular. Desde el punto de vista de los ligandos, quiere decir que su fijación al receptor no necesariamente va a conducir a la producción de una respuesta máxima, sino que existirá, en función de su eficacia y su potencia, una graduación en la respuesta.

La síntesis de nuevos compuestos con capacidad para interaccionar con receptores específicos es el objetivo de la Química Farmacéutica. Todo un arsenal de nuevas metodologías con un importante componente matemático y computacional es empleado para crear modelos tridimensionales de receptores y

ligandos, estudiar sus preferencias conformacionales, dilucidar la naturaleza y magnitud de las fuerzas interatómicas que gobiernan su interacción y analizar el comportamiento dinámico de cada molécula por separado y de sus respectivos complejos. Estos procedimientos ayudan a comprender el comportamiento de los sistemas a nivel submolecular, permiten establecer comparaciones entre teoría y datos experimentales, e incluso permiten realizar predicciones cuantitativas, por lo que constituyen herramientas muy poderosas para diseñar nuevas moléculas con afinidad para el receptor.[6]

2.4 Química computacional en procesos genéticos a nivel molecular.

Desde el descubrimiento de la doble hélice del DNA en 1953 por Watson y Crick, se inició un desarrollo de la biofísica molecular con repercusiones en la genética, medicina y biotecnología. El DNA es el almacén de la información genética, consta de dos cadenas de nucleótidos los cuales pueden ser de cuatro tipos dependiendo de la base nitrogenada que lo componen: Adenina, Guanina, Timina y Citosina. Las bases contienen átomos con características donadoras y aceptoras de protones que pueden formas enlaces de hidrógeno entre sí o con otras moléculas. Aunque existe la posibilidad termodinámica de formar pares entre las diferentes bases; la doble hélice, debido a su estabilidad, está formada por la unión de A-T y G-C.[7] Uno de los primeros estudios de la estructura e interacciones del DNA con otras moléculas ajenas al material genético lo inicia Dickerson[8] utilizando un dodecámero que por su popularidad en el área recibe su nombre (Dickerson-Drew Dodecamer o DDD), donde mediante la estructura cristalina de un oligodesoxinucleótido CGCGAAT-TCGCG y ayudado por rayos X,

detalló la doble hélice del DNA en alta resolución. De esta forma se descubrió que el DDD encontrado era una forma similar a la encontrada por Watson y Crick, sin embargo existen más formas estructurales, determinadas por *n* (número de nucleótidos por vuelta) y *h* (distancia entre unidades repetidas adyacentes). Las variaciones se deben a cambios en la rotación de los grupos alrededor de los enlaces que poseen libertad rotacional (alrededor de seis enlaces en cada monómero). Las ahora llamadas familias estructurales del DNA son tres, principalmente, de acuerdo a su conformación: la Z, B Y A (Figura 1). La forma B representa la estructura general del DNA en las condiciones habituales de las células vivas. Tiene dos surcos que se despliegan a lo largo de la molécula, uno es grande, de 12 Å de amplitud y el otro es pequeño, de 6 Å de amplitud.[9]

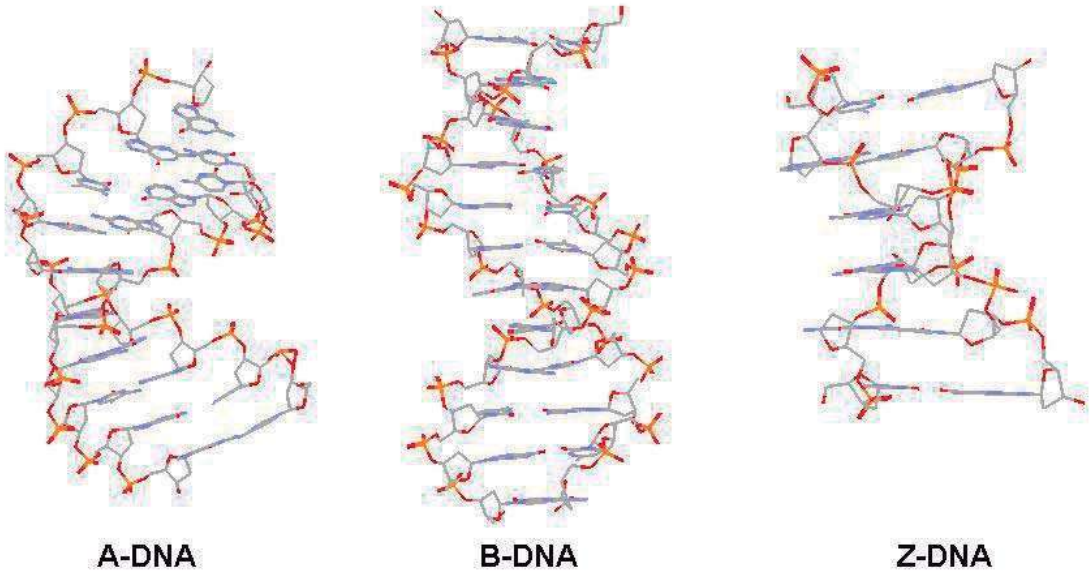


Figura 1. Familias estructurales del DNA. *A-DNA* (*PDB* ID: 213D) B-*DNA* (*PDB* ID: 1BNA) Z-*DNA* (*PDB* ID: 2DCG).

2.5 Reconocedores de surco menor del DNA

Las interacciones fármaco-DNA pueden ser clasificadas dentro de dos grandes categorías: intercaladores y reconocedores de surco. En ambas categorías son posibles los enlaces covalentes y no covalentes. Los intercaladores que, representan la mayoría de interacciones no covalentes con el DNA, son típicamente sistemas poliaromáticos planos que interactúan entre los pares de bases, no presentan mucha especificidad e interrumpen la organización de la doble hélice del DNA. Los reconocedores de surco mayor y menor del DNA difieren en aspectos relevantes en el reconocimiento intermolecular, como las características de los puentes de hidrógeno, disposición estérica, entorno electroestático y microentorno polar. Esto lleva a contrarrestar las preferencias de las moléculas al interaccionar con el DNA. Las proteínas y moléculas grandes prefieren el surco mayor, las moléculas pequeñas se unen a ambos surcos, pero muestran preferencia por ciertas regiones de la doble hélice. Contrario a lo que se pensaría, la preferencia por el surco no está en función de aspectos estéricos sino de interacciones hidrofóbicas o hidrofílicas. Así, algunas regiones peptídicas como los "dedos de zinc" prefieren interactuar con el surco mayor (región hidrofílica) y las moléculas pequeñas poco polares con el surco menor (regiones hidrofóbicas).

Los reconocedores de surco menor tienen estructuras largas y planas que las llevan a adoptar una forma de media luna, conocida como helicoicidad por tomar la forma de la doble hélice, que se ajusta dentro del surco. Las moléculas pequeñas reconocen regiones ricas en A-T (Figura 2), por lo que se dice que son selectivas.[10]

Algunas moléculas y proteínas pequeñas son las que principalmente interaccionan con el surco menor del DNA. Los reconocedores de surco menor como ligandos, a menudo contienen anillos aromáticos simples, tales como pirroles, furanos, bencenos e imidazoles, estos sustituyentes están interconectados por enlaces de libre torsión, lo que permite adoptar la geometría adecuada para la interacción.[11]

La conformación fisiológica del DNA está determinado por cinco tipos de interacciones: uniones covalentes en sus hebras, secuencias de la hebra, puentes de hidrógeno, repulsiones fosfato - fosfato e interacciones con el solvente. [12] El agua es un importante componente estructural del DNA y como su principal solvente, su efecto en la función del DNA es algo muy significativo, especialmente en ambientes celulares donde altas concentraciones de solutos limitan la disponibilidad del agua, por lo que hay que tenerlo siempre en consideración.[13]

Los reconocedores de surco menor han sido utilizados como agentes antibióticos[14], anticancerígenos[15], antimicrobianos de amplio espectro contra enfermedades causadas por protozoarios (tripanosomiasis, leishmaniosis, etc.) e infecciones fúngicas[16], entre otros.

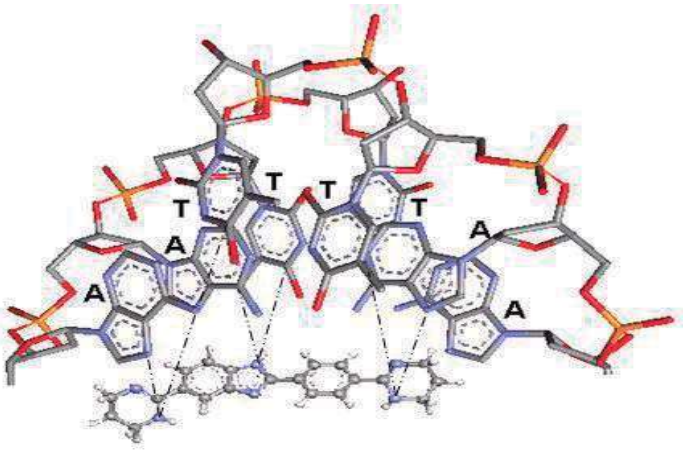


Figura 2. Reconocedores de surco en regiones ricas en A-T.

Es importante señalar que al haber una interacción entre el DNA y el ligando se genera la energía de acomplejamiento (E_c), la cuál está determinada por la ecuación 2.

$$E^{\text{DNA-LIG}} = E^{\text{CPLX}} - [E^{\text{LBC}} + E^{\text{DBC}}]$$

$$\Delta E^{\text{LIGANDO}} = E^{\text{LBC}} - E^{\text{LFC}}$$

$$\Delta E^{\text{DNA}} = E^{\text{DBC}} - E^{\text{DFC}}$$

$$E_c = \Delta E^{\text{DNA}} + \Delta E^{\text{LIGANDO}} + E^{\text{DNA-LIG}} \qquad \textbf{Ec. 2}$$

Donde E^{CPLX} es la energía potencial (absoluta) del complejo final, E^{DBC} es la energía de un solo punto en la conformación que adopta el DNA en el complejo final y E^{LFC} es la energía del ligando en la conformación final. Por su parte, $\Delta E^{\text{LIGANDO}}$ y ΔE^{DNA} son, respectivamente, las energías de distorsión involucradas en el cambio de conformación del ligando y DNA, del estado mínimo libre de la conformación del

farmacóforo. Este parámetro se utiliza para describir la intensidad de afinidad del fármaco por el DNA. La afinidad del ligando hacia el DNA puede ser medida por el logaritmo de la variación de la temperatura de desnaturalización del DNA (log (ΔTm)), cuando interacciona el ligando en relación a su temperatura en estado libre (sin el ligando).

2.6 Temperatura de desnaturalización del DNA o Tm.

Todas las características que le permiten al DNA cumplir con su papel biológico, permiten también manipular los ácidos nucleicos *in vitro* y aislar el segmento de DNA que codifica para una proteína concreta. Como los enlaces que estabilizan a la doble hélice son relativamente débiles pues se trata de puentes de hidrógeno, se pueden romper por calentamiento o por exposición a concentraciones altas de sal. Cuando los puentes de hidrógenos se rompen y por consecuencia se pierde la estructura secundaria original se dice que el DNA está desnaturalizado. Por el contrario, cuando los puentes de hidrógeno están intactos formando la doble cadena del DNA, se dice que el DNA se encuentra en estado nativo. El cambio del estado nativo al desnaturalizado se llama desnaturalización. Si la doble hélice del DNA en estado nativo se somete a calentamiento, los puentes de hidrógeno se rompen y ambas cadenas se separan, entonces, el DNA desnaturalizado se convierte en dos hebras independientes aisladas.[17]

La desnaturalización se lleva a cabo en un intervalo de temperatura muy corto y resulta en cambios muy radicales en cuanto a sus propiedades físicas. Un cambio muy útil es la densidad óptica. Los anillos heterocíclicos de los nucleótidos

absorben luz ultravioleta con un máximo cercano a 260 nm, característico para cada base; sin embargo, la absorbancia difiere a casi el doble cuando el DNA se encuentra como hebra sencilla con respecto a la doble hélice, lo cual permite medir el grado de desnaturalización del DNA conforme se incrementa la temperatura y por ende la absorbancia. El punto medio del intervalo de temperatura en el que se separan las cadenas de DNA se llama temperatura de desnaturalización (Tm: "melting temperatura"). Cuando el DNA está en solución en condiciones más o menos fisiológicas, la Tm está entre 85 - 95 °C. En algunos experimentos, la desnaturalización es comparada con la temperatura de calentamiento del DNA en solución y se construye una gráfica donde la densidad óptica está en función de la temperatura. Estas gráficas son llamadas Curvas de Fusión y tienen forma sigmoidea (Figura 3). En base a la curva de fusión, se determina la desnaturalización total y la Tm siendo esta última la temperatura correspondiente al 50% de desnaturalización.

La Tm está determinada por diversos factores como la concentración de cationes, la presencia de disolventes como el dimetil sulfóxido o formamida, la longitud de la cadena y sobretodo por la proporción de guanina (y citocina) con respecto a adenina (y timina) debido al puente de hidrógeno adicional en el primer par. [18]

La Tm puede ser calculada de manera teórica o experimental mediante el uso de un espectrofotómetro de ultra violeta equipado con variación de temperatura. Otros equipos más sofisticados consideran la entalpía y la entropía, así como parámetros termodinámicos.[19]

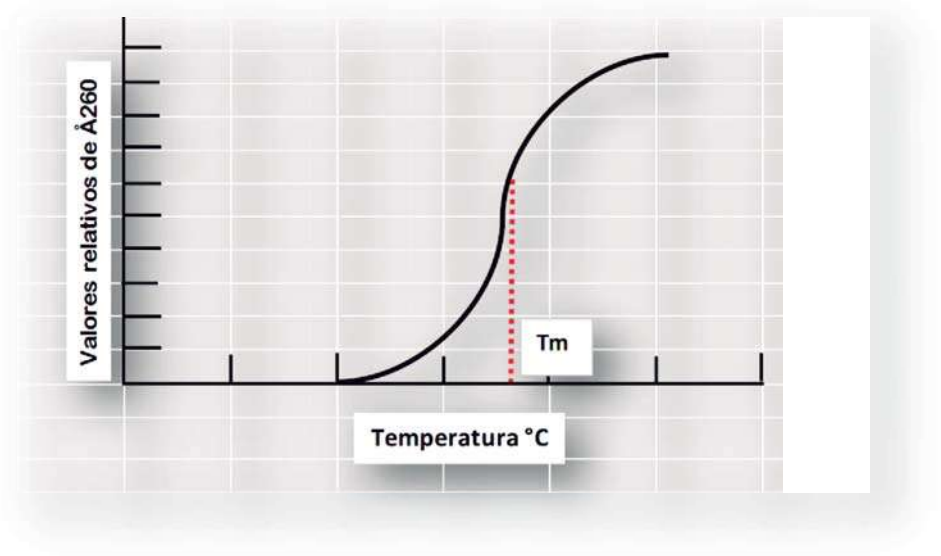


Figura 3. Curva de fusión del DNA.

2.7 Algoritmos Genéticos

La evolución biológica es el proceso de transformación continua de las especies a través de cambios producidos en sucesivas generaciones. Esta evolución da lugar a nuevas especies y permite su adaptación a distintos ambientes. La evolución biológica se produce básicamente por dos procesos: la selección de individuos según sus características y la alteración genética de los cromosomas que almacenan las características de la especie. La selección natural puede ser reproductiva, donde los individuos de ciertas características tienen mayor posibilidad de intervenir en los procesos de reproducción, o puede ser selección ecológica, donde los individuos con ciertas características tienen mayor probabilidad de supervivencia. En ambos casos las características con mayores posibilidades de sobrevivencia son las que favorecen la adaptación al entorno:

mayor capacidad para obtener y procesar el alimento, escapar de los depredadores y otros peligros, resistir a las condiciones ambientales, etc.

La reproducción es la capacidad de toda célula o ser vivo de producir descendientes semejantes a los progenitores y hace posible la continuidad de la vida de la especie.[20] La mutación es todo cambio permanente ocurrido en la secuencia de bases del ADN del organismo y ocurren a nivel molecular.[21] La recombinación o cruzamiento es el intercambio de información genética, por ejemplo, la recombinación entre los cromosomas paternos y maternos de una pareja.[22]

Un algoritmo es por definición una serie de pasos organizados para resolver un problema específico.[23] Los algoritmos genéticos (AG) son métodos de búsqueda basados en los mecanismos de la selección natural y los principios de la genética. Fueron desarrollados tratando de imitar algunos de los procesos observados en la evolución natural. Los mecanismos que guían esta evolución se basan en los principios biológicos siguientes:

1. Los procesos de evolución operan sobre los cromosomas que codifican las estructuras de los seres vivos.
2. La selección natural es la unión entre los cromosomas y la actuación de las estructuras decodificadas. Los procesos de selección natural permiten que se reproduzcan más aquellos cromosomas que codifican estructuras más exitosas.
3. El proceso de reproducción ocurre cuando la evolución toma lugar, bien a través de mutaciones, donde los cromosomas de los hijos difieren ligeramente a los de los padres, o por procesos de recombinación en que los

cromosomas de los hijos varían significativamente respecto a los de los padres mediante la combinación de material genético de los padres.[24]

Al remontarse en el tiempo y hacer un recuento de los personajes que contribuyen a la aparición de los algoritmos genéticos se encuentra que el primero de ellos es el mismo Charles Darwin (1832) y su teoría del origen y evolución de las especies (selección natural). Posteriormente entra en escena Gregor Mendel (1843) con las leyes de la herencia dejando las bases de lo que hoy en día es la genética. Hugo de Vries (1889) introduce el concepto de mutación como explicación a la evolución de las especies. Por último, el creador de los algoritmos genéticos, John Holland (1975) que mediante su tesis doctoral propone los algoritmos genéticos para resolución de problemas, basado en los principios de la herencia y evolución de las especies y finalmente, en 1989, Goldberg populariza su uso.[25]

Un problema de optimización consiste en encontrar dónde se alcanza el máximo o mínimo óptimo de una función real. Este óptimo no tiene por qué ser único y el interés puede estar en encontrar todos los valores de la función donde se alcance el óptimo o sólo uno de ellos; incluso puede ser suficiente con acercarse a la optimización en un alto grado.

Un algoritmo genético, desde el punto de vista de la optimización, es un método poblacional de búsqueda dirigida basada en probabilidad dado que el algoritmo genético emula el comportamiento de una población de individuos que representan soluciones y que evoluciona en base a los principios de la evolución natural: reproducción mediante operadores genéticos y selección de los mejores individuos, correspondiendo éstos a las mejores soluciones del problema a

optimizar. La versión simple del algoritmo genético trabaja siguiendo los siguientes pasos: (Figura 4)

1. Generar una población inicial de soluciones.
2. Seleccionar, de la población actual, las soluciones mejores adaptadas.
3. Cruzar algunas soluciones para obtener su descendencia
4. Mutar algunas soluciones para obtener las soluciones mutadas.
5. Elegir las soluciones que sobreviven y formarán la nueva generación.
6. Si no alcanza el criterio de parada, volver al paso 2.

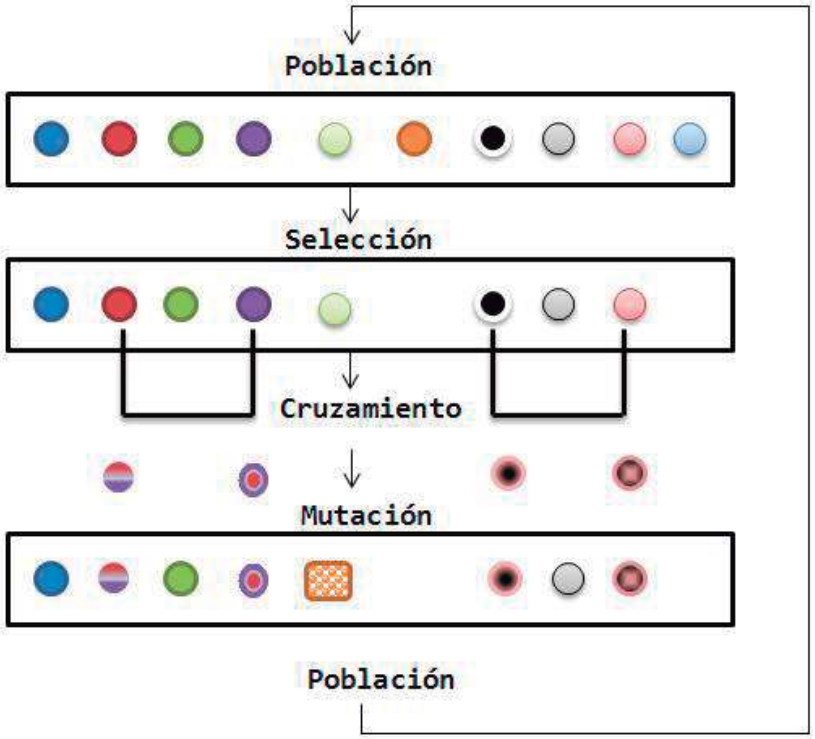


Figura 4: Método general de los algoritmos genéticos.

Al finalizar, la mejor solución de la población es la que se propone como solución del problema.

Los algoritmos genéticos se pueden utilizar para resolver prácticamente cualquier tipo de problema de optimización. Por eso, es necesario establecer la correspondencia entre los distintos elementos del problema y las componentes del algoritmo genético, mostrada en la tabla 1.

Tabla 1. Correspondencia entre problema y algoritmos genéticos.

Evolución Natural	Algoritmo Genético
Evolución	Estrategia
Ambiente	Problema
Población	Conjunto
Individuo	Solución
Adaptación	Calidad
Cromosoma	Representación
Mutación	Movimiento
Cruzamiento	Combinación

La interpretación adecuada de la correspondencia de los elementos es fundamental. La evolución o estrategia guía el proceso al convertir un conjunto arbitrario de soluciones del problema en uno que contenga la solución óptima o una próxima a serlo. El ambiente o problema abordado representa el conjunto de características que condiciona la propia evolución. El objeto de interés es la población de individuos que evoluciona en la naturaleza y los elementos de un

conjunto de soluciones que se van actualizando continuamente. Sus miembros se nombran individuos o soluciones. En ellos se evalúa la adaptación y calidad como soluciones del problema para determinar el papel que juegan en la evolución conjunta y el grado de éxito alcanzado. Los cromosomas y representación son los que almacenan la información necesaria de los individuos para determinar las características de interés. Finalmente, las operaciones que permiten la interacción entre los individuos o soluciones y los cambios que se producen en ellos son las mutaciones o movimiento en el espacio de soluciones y el cruzamiento o combinación de soluciones.[26]

2.8 Estudio de relación cuantitativa estructura-actividad "QSAR"

¿Qué es hoy, el medicamento perfecto? Debe ser una sustancia que, administrada a un paciente cumpla varios criterios: a) que cure la enfermedad, o al menos, sus peores síntomas, b) que sea fácil de administrar al organismo, c) que no sea tóxica, d) que no tenga efectos secundarios, e) que no pierda su efectividad con el uso reiterado. Sin embargo, hasta el momento no hay ninguno que cumpla cabalmente estas características aunque la investigación moderna se está apoyando de varias disciplinas para crear medicamentos más potentes, menos tóxicos, más fácilmente absorbibles o metabolizables y más eficaces. Casi siempre el químico tendrá que hacer muchas, muchísimas moléculas distintas, ensayar una y otra vez hasta obtener alguna que pueda ser benéfica para el uso clínico y que la relación riesgo-beneficio sea apropiada. Los compuestos farmacológicamente activos pueden ser obtenidos de fuentes naturales o de colecciones de compuestos orgánicos sintéticos, pero en cualquier caso lo importante es descubrir un

compuesto que presente una actividad biológica con potencial médico. A este compuesto se le llama cabeza de serie, sin embargo, no tiene por qué ser, y casi nunca será, la molécula final que llegue al mercado. Por ello, es necesario construir nuevas moléculas con estructuras próximas a las cabezas de serie en busca de la máxima potencia terapéutica y los mínimos efectos indeseables. Si las estructuras de los análogos a sintetizar no se alejan demasiado de la del prototipo, los métodos de síntesis y de valoración farmacológica serán también similares y por ende la búsqueda, eficaz y económica. Esto no quiere decir que ante un fármaco prototipo prometedor, la búsqueda del análogo con mejores cualidades haya de seguir efectuándose, en la actualidad, mediante la síntesis de cientos o miles de moléculas relacionadas, para luego ensayarlas de una en una en sus propiedades biológicas. Dada la dificultad y costo en sintetizar todos los análogos posibles, se requiere una metodología más racional para priorizar en su síntesis.

El principal objetivo de los métodos QSAR es justamente predecir la actividad farmacológica de un análogo sin necesidad de prepararlo previamente. Se trata pues, de encontrar una teoría que evitase esa laboriosísima generación de datos. La teoría se apoya en una idea básica: que todas las propiedades de una sustancia, sean físicas, químicas o biológicas, están en función de su estructura molecular. Cualquier modificación en ésta conlleva una variación de propiedades, en mayor o menor grado. Esto es en lo que refiere al aspecto cualitativo. El problema que se plantea ahora es cuantificar esa dependencia. ¿En qué medida se modifica la potencia del fármaco si una subunidad de su estructura molecular se sustituye por otro fragmento? Ciertamente, se puede cuantificar mediante una ecuación, estadísticamente significativa que correlacione las propiedades moleculares de una serie de compuestos activos con la actividad experimental que presentan. Se

requiere por lo tanto, una serie de datos de actividad sobre una pequeña familia seleccionada de compuestos y se efectúa la interpolación o extrapolación para predecir las propiedades de nuevos análogos aún no sintetizados. Las ventajas son obvias: permite que el químico sintético dedique sus esfuerzos a obtener análogos que deberían tener mejor actividad, sin perder el tiempo en otros que se predicen innecesarios, además que si se obtiene algún análogo que no cumple la ecuación, se sabe que hay alguna propiedad más que debe ser considerada para refinar el cálculo.[39] Existen muchos propósitos prácticos de un estudio QSAR y esta técnica es utilizada extensamente en muchas situaciones. Los principales propósitos son:

- Predecir la actividad biológica y las propiedades físicoquímicas por métodos racionales.
- Comprender y racionalizar los mecanismos de acción dentro de diversas series de compuestos.

Fundamentalmente el desarrollo de estos modelos tiene como objetivo:

- Ahorrar en el costo al desarrollar nuevos productos.
- La predicción puede remplazar, reducir e incluso eliminar la necesidad de largas y costosas pruebas con animales y obviamente el sufrimiento que conlleva.
- Otras áreas lo promueven, como la Química Ecológica, al aumentar la eficiencia y eliminar pérdidas.[27]

Dentro de las aplicaciones que se le ha dado a los estudios QSAR, existe una lista muy extensa: nuevos fármacos antivirales, [28] compuestos anticancerígenos

inorgánicos, [29] fármacos antituberculosos, [30] compuestos anti-*Trypanosoma cruzi*[31] y herbicidas, [32] solo por citar algunos.

Los métodos QSAR han agrupado históricamente a todas las técnicas que han intentado establecer modelos empíricos de comportamiento sobre familias de compuestos biológicamente activos para obtener óptimos de actividad a partir de los datos de comportamiento de un número limitado de productos. La primera vez que se habló de un modelo cuantitativo de la relación estructura-actividad fue por Bruice, Karasch y Winzler en 1956. Después vinieron Free y Wilson en 1964. El modelo de Free-Wilson está basado en el uso de series de compuestos derivados de una estructura común, en los que se observan las alteraciones producidas en su actividad biológica en función de la presencia de sustituyentes diversos.

Simultáneamente en 1964, Corwin Hansch describe un trabajo similar, donde correlaciona la actividad biológica de series de productos con sus propiedades físicoquímicas. El modelo extratermodinámico de Hansch brinda una explicación matemática de la interacción del fármaco con su receptor en la ecuación:

$$\ln A = f_h\,(X_h) + f_e(X_e) + f_s\,(X_s) + Cte \qquad \textbf{Ec. 3}$$

Donde A es la actividad y f_h, f_e, f_s son funciones de índices o parámetros hidrofóbicos, electrónicos o estéricos respectivamente. El término extratermodinámico surge porque las relaciones se describen en términos termodinámicos, aunque no se deducen de sus leyes.[33]

La metodología QSAR sigue, independientemente del modelo que se utilice, algunos pasos comunes que se muestran en la figura 5. El punto de partida en todos los casos implica la existencia de un prototipo cabeza de serie, que se define como un producto que muestra actividad en relación con el objeto terapéutico buscado.

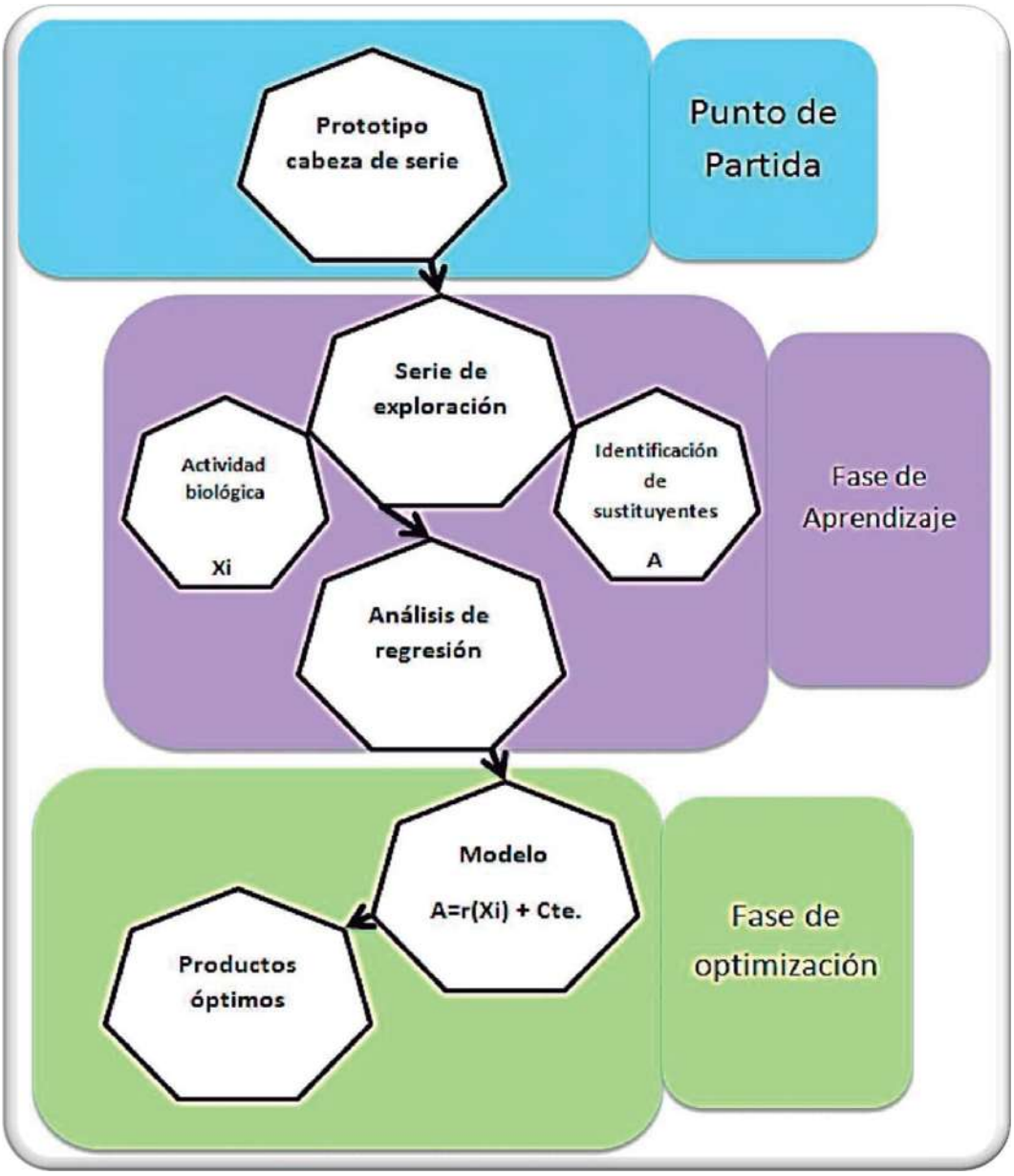


Figura 5. Metodología QSAR

Cuando se dispone de un prototipo, es preciso diseñar una serie de exploración, que está constituida por un conjunto de productos análogos del prototipo, que permitan el establecimiento de las primeras relaciones estructura-actividad. Los miembros de la serie de exploración tienen que ser sintetizados y analizados en su actividad. Los compuestos o los sustituyentes presentes en ellos, han de ser identificados por descriptores, que serán utilizados como variables independientes (Xi) en el modelo. Por otro lado, los valores de actividad biológica se utilizan como variable dependiente (A) en el modelo.

Ya establecidos los conjuntos Xi y A, se utilizan técnicas de regresión múltiple para obtener un modelo donde la actividad está expresada en forma de la sumatoria de una constante y de las contribuciones de los sustituyentes variables utilizados. El modelo se analiza estadísticamente para evaluar su capacidad descriptiva y de predicción. Cuanta mayor calidad estadística tenga el modelo, más fiables serán las predicciones de actividad obtenidas. La fase de optimización implica que una vez obtenido un modelo de buena calidad, sea posible calcular los productos con actividad óptima dentro de la familia estudiada [34].

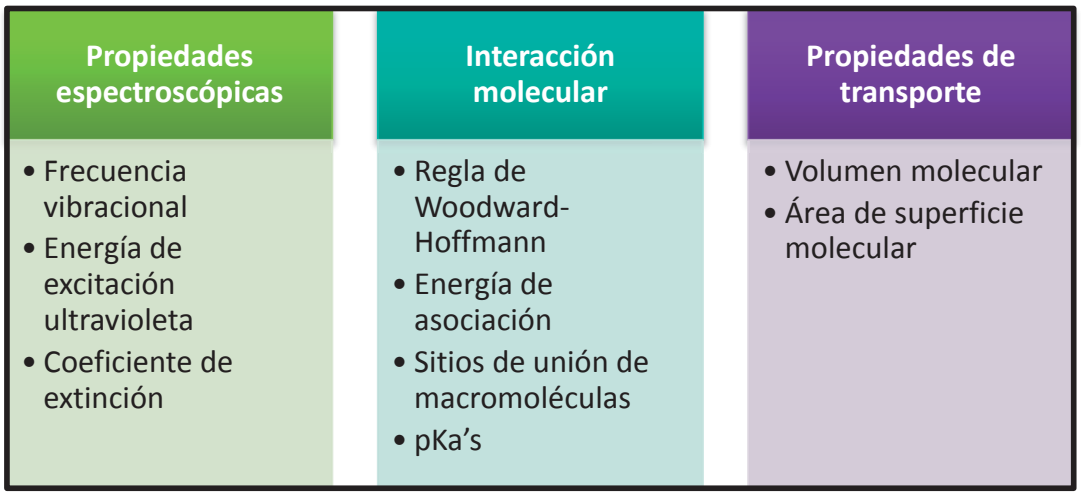
Hay tres principales partes del QSAR que se aprovechan en la búsqueda científica: el concepto de estructura molecular, la definición de los descriptores moleculares y las herramientas quimioinformáticas. El concepto de estructura molecular es representado por su descriptor molecular teórico que tiene una estrecha relación con las propiedades experimentales de las moléculas, pues de estas propiedades experimentales surgen los descriptores moleculares teóricos.

2.9 Descriptores Moleculares

En décadas pasadas, muchos científicos buscaron centrar su atención en cómo capturar y convertir, de manera teórica, la información codificada en la estructura molecular de alguna manera o usar números para establecer relaciones cuantitativas entre estructura y sus propiedades, actividad biológica u otras propiedades experimentales. Estas propiedades están implícitamente dentro de la estructura molecular. En la tabla 2 se muestran ejemplos de estas propiedades.

Tabla 2. Propiedades moleculares implícitas.

Datos geométricos	Termodinámica y energía	Propiedades electrónicas
• Longitud de enlace • Ángulo de enlace • Ángulo de torsión • Estructura tridimensional • Distancia interatómica • Intermediarios de reacción • Estados de transición	• Energía molecular • Calor de formación • Población conformacional • Entropía • Energía de activación • Superficie de energía potencial • Camino de reacción • Energía de solvatación	• Distribución de carga • Momento dipolar • Potencial de ionización • Afinidad electrónica • Afinidad protónica • Polarización • Campo de potencial electrostática



Los descriptores moleculares son representaciones matemáticas formales de una molécula, obtenidos mediante un algoritmo específico y aplicado para definir una representación molecular de un procedimiento experimental específico: son el resultado final de un procedimiento lógico y matemático que transforma la información química codificada dentro de una representación simbólica de una molécula en un conveniente número como resultados de algunos experimentos estandarizados.

Los descriptores moleculares juegan un papel fundamental en el desarrollo de modelos para la química, ciencias farmacéuticas, protección ambiental, toxicología, ecotoxicología, investigaciones en salud y control de calidad. Su interés en la comunidad científica se refleja en la gran cantidad de descriptores que han sido propuestos: más de 5000 de ellos derivados de diferentes teorías y enfoques son definidos y computarizados usando software de simulación de la estructura química.

Los descriptores moleculares se dividen en dos principales grupos: los que requieren de mediciones experimentales, tales como momento dipolar, polarizabilidad y en general, propiedades físico-químicas, y descriptores moleculares teóricos, que son derivados de una representación simbólica de las moléculas y puede ser, además, clasificados de acuerdo a los diferentes tipos de representación molecular. La diferencia fundamental entre los descriptores teóricos y los experimentales es que los descriptores teóricos no contienen errores estadísticos debido al error de las mediciones experimentales. Además, las necesidades de facilitar el cálculo y la aproximación numérica son por sí mismos asociados con un error inherente, aunque en la mayoría de casos la dirección, pero no la magnitud del error es conocido. Por otra parte, dentro de una serie de componentes relacionados, el error es usualmente considerado como constante. Todos los tipos de error están ausentes hasta en el más simple descriptor teórico, ya que las características estructurales de los descriptores son derivados directamente de teorías matemáticas exactas.

2.10 Descriptor D_{CL}

Las necesidades de adoptar nuevos descriptores para estudios de interacciones intermoleculares específicas han aumentado conforme se utilizan las técnicas computacionales en la investigación científica. De esta forma surgió en nuestro grupo de trabajo un nuevo descriptor molecular arbitrariamente denominado D_{CL}, [35] relacionado con el reconocimiento topológico del DNA. Este descriptor fue obtenido a partir de la distancia de la molécula geométricamente optimizada y la topología del ADN asumiendo que se requiere una distancia

específica para un reconocimiento óptimo entre ambas moléculas. Cuantitativamente fue obtenido de la siguiente manera: la distancia entre los grupos amino, responsables de la interacción con el DNA por puentes de hidrógeno, dividida entre 4.0202 (o 3.4/cos 36), que representa el número de pares de bases más uno que el ligante de surco debe reconocer, por lo tanto, si se resta el número entero más cercano al valor obtenido, el valor absoluto de la diferencia resultante será la separación relativa de reconocimiento y abarcará un intervalo de 0 a 0.5 siendo la interacción "más óptima" cero y la "menos óptima" de 0.5. Lo anterior puede ser apreciado en la figura 6 y en la ecuación 4.

$$\mathbf{D_{CL}} = |\ (\text{Distancia entre } NH_2 - NH_2/4.202) - N| \qquad \textbf{Ec. 4}$$

En donde N es el número entero más cercano.

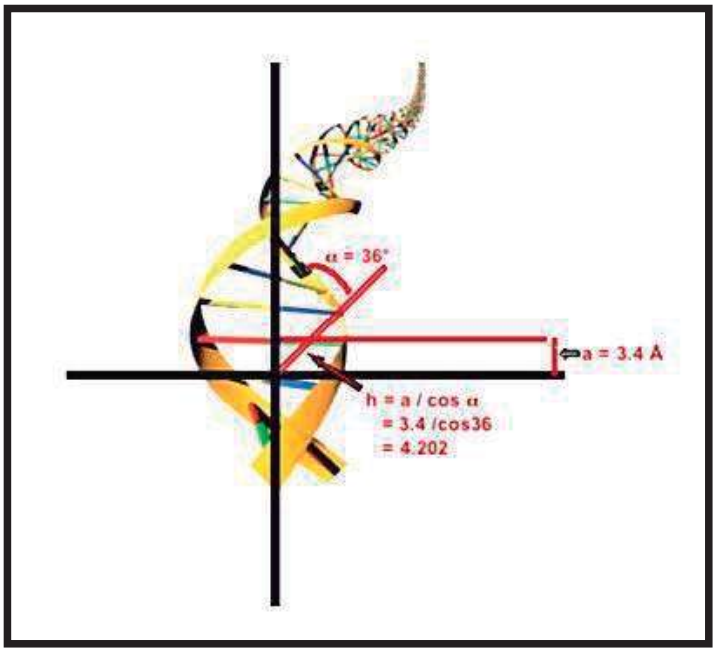


Figura 6. Representación esquemática del descriptor D_{CL}

Dado lo anterior, resulta interesante explorar si el descriptor D_{CL} tiene aplicación a compuestos diferentes a los estudiados en la publicación original (únicamente alcanodiamidas), sobre todo con actividad biológica demostrada en DNA (ΔTm) y como consecuencia aprovecharlo para buscar compuestos novedosos con mayor actividad y menor toxicidad.

2.11 Docking

A medida que se fueron cristalizando y analizando complejos macromolécula-ligando se fueron estableciendo las características generales de la unión, se puso de manifiesto que el ligando se mantiene en una posición específica, dentro de la diana molecular, anclado por muchas interacciones con diferentes grupos del receptor. Una función importante de la estructura macromolecular en su conjunto es mantener estos grupos en la orientación relativa adecuada para definir las características de la cavidad y constituir un sitio de unión específico. Las contribuciones individuales de cada grupo a la energía total de unión pueden ser débiles, pero la suma de todas ellas puede hacer que el ligando se una muy firmemente. En ausencia de datos cristalográficos sobre el modo preciso de unión del ligando a la macromolécula, es preciso explorar las diversas posibilidades de interacción entre ambas moléculas. Para que se forme un complejo, es requisito que la energía de interacción (ΔE), sea negativa, de acuerdo con la ecuación 5:

$$\Delta E = E_{(ligando\text{-}receptor)} - E_{ligando} - E_{receptor} \qquad \textbf{Ec. 5}$$

La computación de ΔE implica conocer tanto las conformaciones óptimas de ligando y receptor por separado como la conformación más estable del complejo. La búsqueda del complejo óptimo no es trivial en la mayoría de las ocasiones, y su modelado implica tener en consideración un buen número de factores, entre los que se encuentran:

a) Conocimiento del sitio de unión en la macromolécula.
b) Determinación de la orientación relativa del ligando con respecto al sitio receptor, pues los casos de múltiples modos de unión están suficientemente documentados. Los métodos teóricos normalmente buscan aquella orientación que da lugar a la energía de interacción más favorable, aunque cuando se ha podido comparar con resultados cristalográficos, se ha llegado a la conclusión de que las orientaciones calculadas como más favorables no tienen por qué coincidir necesariamente con los observados experimentalmente.
c) Caracterización de la conformación óptima del ligando en el sitio de unión.
d) Evaluación de los posibles cambios conformacionales en el sitio receptor como consecuencia de la unión.

Las técnicas manuales de modelado del complejo ligando-receptor tratan inicialmente al sitio diana como si fuera completamente rígido, mientras que la conformación del ligando se va ajustando de forma interactiva, tanto por movimientos de traslación y rotación como mediante giros de enlaces rotables. Esta maniobra de acoplamiento ("docking", por analogía con la atracada de los

buques en puerto) se simplifica considerablemente si se cuenta con información visualmente comprensible sobre las características químicas de ambas moléculas. Los cálculos energéticos proporcionan asimismo una medida de la energía de interacción y sirve como guía para elegir la orientación preferida de una molécula respecto de otra.

Dada la necesidad de búsqueda de nuevos compuestos en enfermedades tan relevantes como el cáncer, en el presente estudio se establece un estudio QSAR de una serie de reconocedores de surco del DNA mediante algoritmos genéticos utilizando 1665 descriptores moleculares teóricos dentro de los cuales se incluye al descriptor D_{CL}. Los resultados son apoyados por estudios de anclaje molecular "Docking".

3. HIPÓTESIS

El descriptor D_{CL} representa una herramienta importante en el reconocimiento topológico entre el DNA y un ligando y puede ser incorporado en modelos estadísticamente significativos que expliquen la actividad de reconocedores de surco del DNA.

4. OBJETIVOS

General

Realizar un estudio de relación cuantitativa estructura-actividad (QSAR) para obtener información geométrica que ayude al acoplamiento de los reconocedores de surco en el DNA mediante una serie de datos descritos en la literatura introduciendo al descriptor D_{CL}.

Particulares

- Obtener modelos estadísticamente significativos que incluyan al descriptor D_{CL}.
- Predecir actividad biológica (log (ATm)) de un péptido propuesto utilizando modelos que incluyan al descriptor D_{CL}.
- Comprobar la interacción del péptido con el DNA mediante un estudio Docking.

5. METODOLOGÍA

Se realizó el estudio QSAR utilizando una serie de exploración de 27 bisamidinas aromáticas 1 - 27 (Figura 7), probadas experimentalmente como reconocedoras de surco del DNA complementarias al dodecámero de Dickerson Drew (DDD) con respecto a la ΔTm que estas inducen en el DNA, acorde a una base de datos descrita en la literatura[36]. Se modelaron las estructuras con el programa HyperChem[37] y con el mismo, se optimizaron geométricamente con el método semiempírico AM1, utilizando como base el algoritmo Polak-Ribiere[38], con parámetros de condiciones de terminación de 0.1 kcal/ (Å mol) o 465 como máximo. Posteriormente se procedió a obtener los descriptores moleculares mediante el programa DRAGON[39], donde se calcularon un total de 1664 descriptores para cada molécula, agrupados en 20 familias. Adicionalmente, se incorporó el descriptor D_{CL}, que se obtuvo de acuerdo a la literatura[35] teniendo un total de 1665 descriptores moleculares.

Se realizó el estudio QSAR utilizando una serie de exploración de 27 bisamidinas aromáticas 1 - 27 (Figura 7), probadas experimentalmente como reconocedoras de surco del DNA complementarias al dodecámero de Dickerson Drew (DDD) con respecto a la ΔTm que estas inducen en el DNA, acorde a una base de datos descrita en la literatura[36]. Se modelaron las estructuras con el programa HyperChem[37] y con el mismo, se optimizaron geométricamente con el método semiempírico AM1, utilizando como base el algoritmo Polak-Ribiere[38], con parámetros de condiciones de terminación de 0.1 kcal/ (Å mol) o 465 como máximo. Posteriormente se procedió a obtener los descriptores moleculares mediante el programa DRAGON[39], donde se calcularon un total de 1664

descriptores para cada molécula, agrupados en 20 familias. Adicionalmente, se incorporó el descriptor D_{CL}, que se obtuvo de acuerdo a la literatura[35] teniendo un total de 1665 descriptores moleculares.

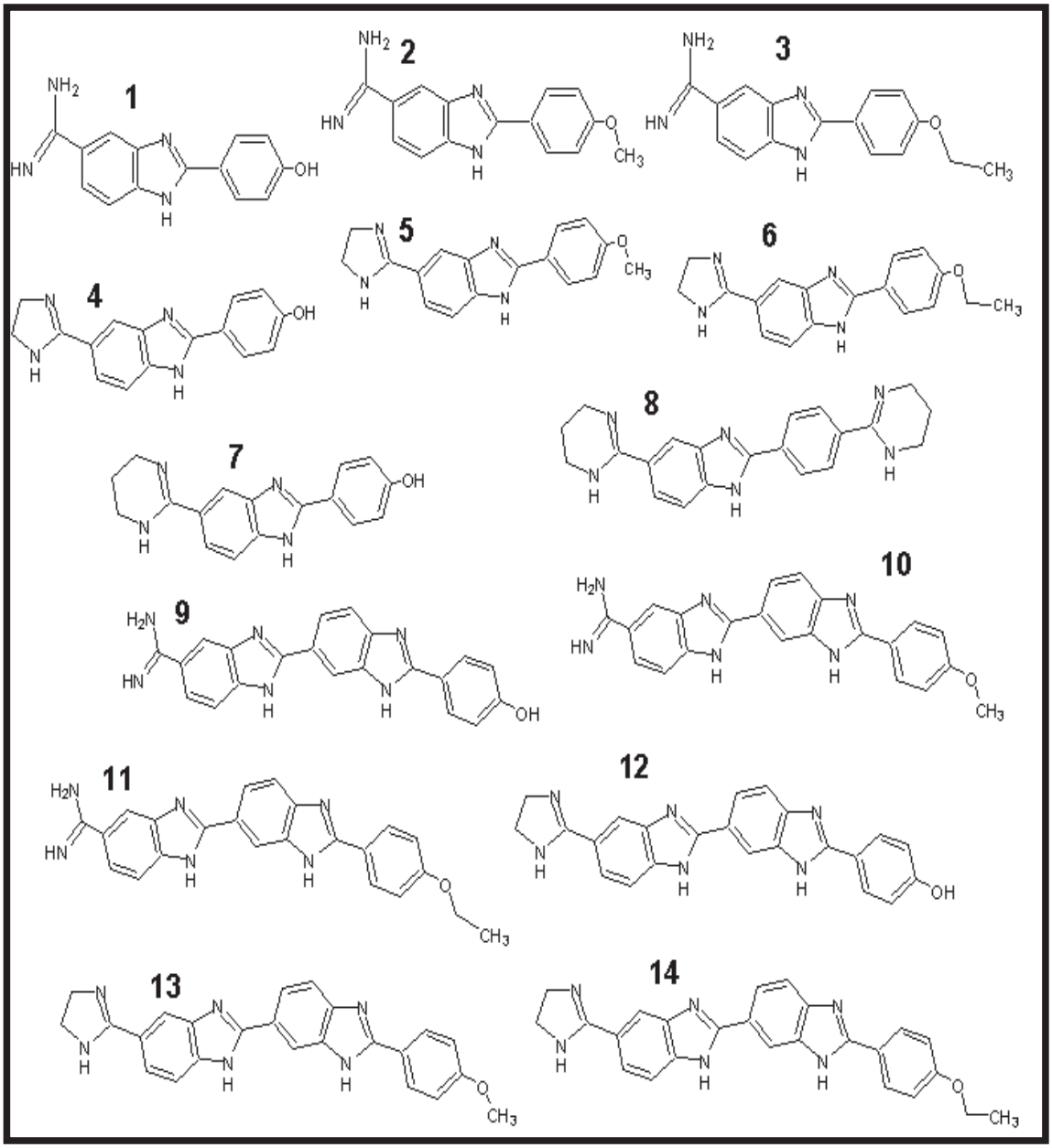


Figura 7. Serie de exploración.

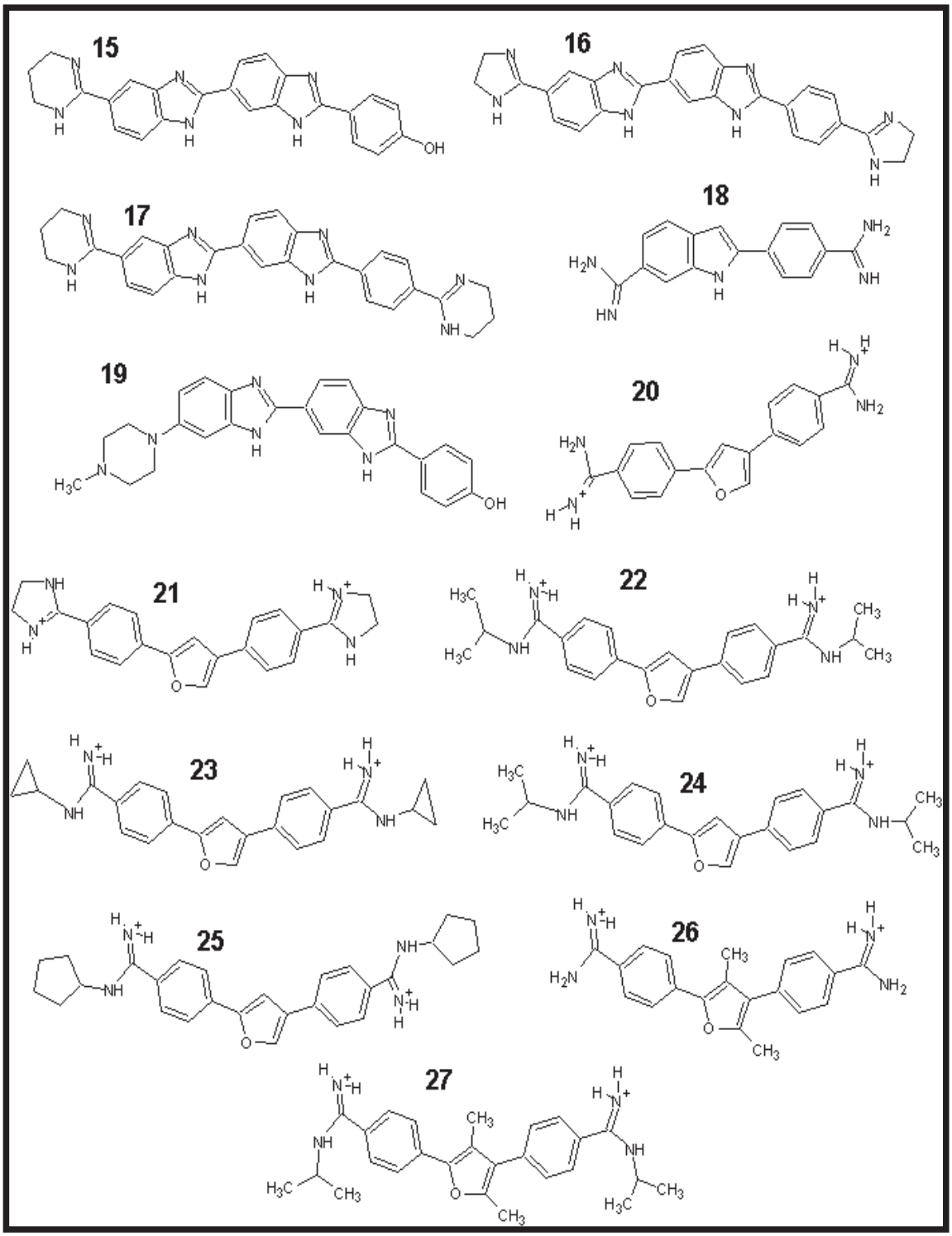


Figura 7. Serie de exploración (*cont.*).

Los descriptores fueron importados al programa MobyDigs.[40] Este software trabaja siguiendo el modelo de los Algoritmos Genéticos y se consideran como parámetros el tamaño de población (100), cuatro variables máximas permitidas en un modelo, mutación máxima de 0.3, nueva generación en cada población cada 3000 ciclos e incrementar variables cada 30000 ciclos. La variable independiente se tomó como -log (1/y) en donde "y" es igual al log (ΔTm). La evolución se lleva a cabo hasta un máximo de 700 evoluciones o bien, hasta que los cincuenta mejores modelos se encuentren estables. Este programa solamente respalda diez familias, por lo que se procedió de la misma forma dos veces, primero de la familia 1 a la 10 y después de la 11 a la 20. Cuando llegan al criterio de parada, los 50 mejores modelos de cada parte se unen y se evolucionan juntos para presentar resultados más homogéneos y confiables. El criterio de parada de esta parte es de más de 1200 evoluciones o hasta que los primeros cincuenta modelos se mantuvieron estables.

6. RESULTADOS Y DISCUSIÓN

Al finalizar el cálculo en MobyDigs, se obtuvieron los mejores cincuenta modelos que se muestran en la tabla 3, entre los cuales se encontró que los modelos 39 y 48 consideran el descriptor D_{CL} con buena significancia estadística.

Tabla 3. Resumen de los cincuenta "mejores" modelos obtenidos a partir del cálculo por algoritmos genéticos. Se muestran sus descriptores y principales datos estadísticos.

ID	Descriptores Involucrados	R^2	Q^2	Q^2boot	Kx	F	s
1	D/D ATS1m EEig02r GGI9 ***	95.24	92.82	91.34	71.9	104.94	0.026
2	Xu ATS1m EEig02r GGI9***	95.51	92.66	91.59	77	111.63	0.026
3	ATS2m MATS8e MATS8p GATS1p	95.35	92.44	91.86	23.35	107.63	0.026
4	EEig02r GGI4 GGI5 GGI8***	95.32	92.41	90.89	60.05	106.85	0.026
5	EEig02r GGI5 GGI8 LP1**	95.27	92.36	89.78	56.7	105.79	0.026
6	piPC01 MATS8e MATS8p GATS1p	95.13	92.34	91.6	21.3	102.51	0.027
7	ATS1m EEig02r GGI9 AEigm***	95	92.04	90.89	74.1	99.8	0.027
8	ATS1m EEig02r GGI9 AEigZ***	95	92.04	90.88	74.1	99.8	0.027
9	EEig02r GGI5 GGI8 H-048*^	95.14	92.02	90.06	52.12	102.67	0.027
10	EEig02r GGI5 GGI8 C-033*^	95.14	92.02	90.05	52.12	102.67	0.027
11	ATS1m EEig02r GGI9 Eig1Z***	94.99	92.01	90.83	74.12	99.56	0.027
12	EEig02r ESpm09u GGI5 GGI8***	95.03	91.94	90.48	63.36	100.36	0.027
13	EEig02r ESpm10u GGI5 GGI8***	95.05	91.94	90.41	63	100.81	0.027
14	IVDM MATS8v MATS8e GATS1p	94.74	91.83	91.25	22.2	94.47	0.028
15	EEig02r ESpm07u GGI5 GGI8***	94.86	91.79	90.56	64.3	96.91	0.027
16	EEig02r GGI5 GGI8 E1e*^	95.12	91.79	90.85	48.08	102.41	0.027
17	EEig02r GGI5 GGI8 VEA1***	94.98	91.75	90.82	63.55	99.39	0.027
18	EEig02r ESpm08u GGI5 GGI8***	94.84	91.74	90.63	64.19	96.54	0.027
19	EEig02r GGI3 GGI5 GGI8***	94.97	91.71	88.91	60.94	99.18	0.027
20	ATS1m MATS8v MATS8e GATS1p	94.56	91.65	90.89	22.6	91.19	0.028
21	ATS1m MATS8e MATS8p GATS1p	94.72	91.64	91.02	24.71	94.12	0.028
22	EEig02r GGI5 GGI8 C-026***	94.8	91.6	90.45	55.22	95.66	0.028
23	EEig02r GGI5 GGI8 C-040***	94.8	91.6	90.38	55.22	95.66	0.028
24	EEig02r GGI5 GGI8 E1u**	95.04	91.6	90.52	49.75	100.54	0.027
25	EEig02r GGI5 GGI8 nCconj**	94.81	91.54	90.56	55.36	95.96	0.027
26	IDDE EEig02r GGI5 GGI8**	94.94	91.52	89.78	55.89	98.45	0.027
27	EEig02r GGI5 GGI8*^	94.41	91.51	90.92	57.79	123.88	0.028
28	EEig02r GGI5 GGI8 JGI3*^	94.8	91.48	89.12	52.95	95.62	0.028
29	piPC01 MATS8v MATS8e GATS1p	94.43	91.42	90.49	19.17	89.06	0.028
30	EEig02r GGI5 GGI8 AEigZ***	94.68	91.39	90.49	66.29	93.35	0.028
31	EEig02r GGI5 GGI8 AEigm***	94.68	91.39	90.44	66.29	93.35	0.028

32	EEig02r GGI5 GGI8 Eig1Z***	94.68	91.39	90.55	66.28	93.36	0.028
33	MATS8p EEig02r GGI5 GGI8*^	94.56	91.39	90.24	48.68	91.18	0.028
34	EEig13x EEig02r GGI5 GGI8***	94.51	91.38	89.96	64.17	90.41	0.028
35	X0Av EEig02r GGI5 GGI8***	94.68	91.37	90.36	60.53	93.4	0.028
36	VDA EEig02r GGI5 GGI8***	94.67	91.36	90.38	65.04	93.28	0.028
37	nC EEig02r GGI5 GGI8***	94.56	91.36	90.44	66.71	91.18	0.028
38	D/D EEig02r GGI5 GGI8***	94.63	91.31	90.4	68.6	92.5	0.028
39	**G(N..N) RDF095u E1e D_{CL} ***	**93.26**	**91.31**	**85.23**	**31.66**	**72.61**	**0.031**
40	RBN EEig02r GGI5 GGI8***	94.56	91.31	89.95	60.8	91.19	0.028
41	GATS1p EEig02r GGI5 GGI8*^	94.55	91.3	90.17	41.9	91.05	0.028
42	X1v EEig02r GGI5 GGI8***	94.53	91.29	90.27	63.94	90.65	0.028
43	IDDM TIC0 GATS7v*^	92.78	89.33	89.09	59.99	94.31	0.032
44	IAC IDDM GATS7v*^	92.78	89.33	89.03	59.99	94.31	0.032
45	GATS7v G(N..N) Ui*	87.02	83.25	82.4	46.63	49.15	0.043
46	GATS7v EEig02r	86.42	82.3	82.4	30.07	73.15	0.043
47	E1u nCbH C-002 Ui**	89.15	81.18	79.48	58.69	43.14	0.04
48	**G(N..N) D_{CL}**	**83.49**	**80.98**	**66.34**	**22.88**	**58.15**	**0.047**
49	GATS7v Ui	84.58	80.31	79.94	28.79	63.07	0.045
50	nCar C-033 Ui*^	88.1	79.74	77.28	51.59	38.82	0.047

*** Modelo que presentan tres o más combinaciones de descriptores con alta correlación. **Modelo que presenta dos combinaciones de descriptores con alta correlación. *^Modelo que presenta una combinación de descriptores con alta correlación. No apto para fragmentación. *Modelo que presenta una combinación de descriptores con alta correlación. Apto para fragmentación.

Todos los modelos fueron estadísticamente validados, de forma interna y externa. La validación interna se realizó mediante los datos que proporciona el mismo programa MobyDigs, entre ellos la Q^2, R^2, Q^2boot, F, s y Kx.

Así mismo, mediante el programa Excel, [41] se calculó la correlación de Pearson (Tabla 4) entre cada combinación de dos descriptores que presentan los modelos y se descartaron los modelos que sobrepasan 0.5 de correlación entre sus descriptores, quedando solamente 8 modelos adecuados y uno más que puede ser sometido a fragmentación.

Tabla 4. Correlación de Pearson entre los descriptores del modelo 39.

Descriptor	G(N..N)	RDF095u	E1e	D_{CL}
G(N..N)	1			
RDF095u	0.45090029	1		
E1e	0.10532029	0.53286131	1	
D_{CL}	-0.24402103	-0.15681401	0.07324901	1

También se calculó la estimación lineal y siguiendo la ecuación base (Ecuación 6) se obtienen los valores que conforman los modelos completos de acuerdo a la ecuación general (Ecuación 6).

$$y = mx + b \quad \text{o bien} \quad y = m1x1 + m2x2 + \cdots + b \qquad \textbf{Ec. 6}$$

Adicionalmente y con la finalidad de valorar la capacidad predictiva del modelo, se extrajo de la serie de datos el compuesto 8 y se calculó su actividad con los modelos obtenidos para cada compuesto y para el compuesto 8. Los datos calculados en las ecuaciones 7, 8 y 9 se compararon con los datos experimentales (Tabla 5).

Tabla 5. Actividad calculada con las tres ecuaciones de los modelos obtenidos.

	Valores de descriptores				log(ΔTm)	log(ΔTm) calculado		
Mol	**G(N..N)**	**E1e**	**D_{CL}**	**RDF095u**	**Experimental**	**Ecuación 7**	**Ecuación 8**	**Ecuación 9**
1	24.32	0.58	0.49	4.76	0.845	0.797	0.799	0.783
2	23.89	0.56	0.41	7.73	0.792	0.848	0.849	0.814
3	24.32	0.55	0.5	7.06	0.69	0.79	0.791	0.743
4	24.44	0.57	0.48	4.63	0.748	0.804	0.807	0.789

5	24.51	0.56	0.47	6.47	0.875	0.81	0.812	0.768
6	24.57	0.56	0.46	2.24	0.845	0.813	0.818	0.77
7	24.12	0.58	0.46	6.57	0.845	0.812	0.814	0.808
8	105.1	0.6	0.09	12.54	1.233	1.352	1.352	1.405
9	95.08	0.58	0.01	11.62	1.362	1.399	1.4	1.4
10	95.19	0.58	0.14	8.7	1.362	1.263	1.267	1.267
11	95.17	0.57	0.01	13.05	1.362	1.399	1.398	1.369
12	95.61	0.58	0.03	11.05	1.378	1.382	1.384	1.394
13	95.45	0.58	0.02	16.14	1.387	1.391	1.386	1.408
14	95.6	0.57	0.07	11.55	1.412	1.339	1.34	1.323
15	94.91	0.59	0.15	12.86	1.405	1.25	1.25	1.266
16	250.1	0.58	0.43	15.27	1.479	1.54	1.543	1.482
17	248.2	0.6	0.46	17.58	1.487	1.487	1.487	1.477
18	78.29	0.57	0.24	5.17	1.233	1.118	1.124	1.083
19	98.5	0.58	0.07	11.24	1.253	1.346	1.347	1.356
20	59.42	0.58	0.35	7.94	0.959	0.976	0.978	0.978
21	59.68	0.58	0.33	8.02	0.857	0.994	0.996	0.998
22	59.32	0.62	0.36	12.94	1.037	0.969	0.966	1.048
23	59.06	0.66	0.48	15.29	1.037	0.884	0.88	1.024
24	58.75	0.61	0.5	12.92	0.968	0.868	0.866	0.913
25	59.09	0.68	0.38	22.59	1.113	0.951	0.941	1.157
26	59.72	0.52	0.35	8.18	0.813	0.974	0.975	0.862
27	59.31	0.57	0.39	25.01	0.732	0.946	0.934	0.911

Se observó que el modelo 39 sobrepasa el 0.5 en la correlación de Pearson, establecido como límite entre los descriptores E1e y RDF095u, sin embargo, se puede hacer una fragmentación, cuando la correlación no sobrepasa en gran medida el 0.5 establecido como límite, para explorar modelos donde se descarten estos descriptores y ponderar su presencia en el modelo. Así, obtenemos como resultado un total de tres ecuaciones: la ecuación 7 que deriva del modelo 48 y dos más, la ecuación 8 y 9, que surgen de la fragmentación del modelo 39:

$$\log(\Delta Tm) = 0.0012\, G(N..N) - 0.3368\, D_{CL} + 0.0371 \qquad \textbf{Ec. 7}$$

n= 27 R^2 = 0.8349 F=58.1512 s=0.0469 Q^2= 80.98

$$\log(\Delta Tm) = 0.0012\, G(N..N) - 0.0004\, RDF095u - 0.3374\, D_{CL} + 0.0403 \qquad \textbf{Ec. 8}$$

n= 27 R^2 = 0.8352 F=37.1585 s=0.0479 Q^2= 80.79

$$\log(\Delta Tm) = 0.0011\, G(N..N) + 0.8935\, E1e - 0.3567\, D_{CL} - 0.4731 \qquad \textbf{Ec. 9}$$

n= 27 R^2 = 0.8970 F=63.8463 s=0.0379 Q^2= 87.57

6.1 Interpretación de los descriptores moleculares involucrados en los modelos encontrados.

La interpretación de los descriptores puede resultar bastante difícil o en ocasiones abstracta pues cabe recordar que se trata en su mayoría de descriptores teóricos. A continuación se analizará cada uno de los descriptores para facilitar su posible interpretación física y relación con los cambios de actividad.

G(N..N). Descriptor geométrico que considera la suma de las distancias geométricas entre Nitrógenos. No involucra la forma o geometría helicoidal del DNA y la distancia entre nitrógenos se obtiene a partir de la estructura optimizada. La contribución dentro del modelo es justificable gracias a que la formación del complejo DNA-ligando requiere un adecuado reconocimiento geométrico.

RDF095u. Función de distribución radial - 9.5/ sin ponderar. Forma parte de los descriptores de distribución radial (*Radial Distribution function*). Fueron propuestos en base a las diferentes funciones de distribución radial que son usados comúnmente para calcular transformaciones moleculares. Formalmente, la función de distribución radial de un conjunto de átomos *A* puede ser interpretada como la probable distribución de encontrar un átomo en un volumen esférico de radio *R*. La forma general de la función de distribución radial está representada en la ecuación 10.

$$g(R) = \mathrm{f} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} wi \cdot wj \cdot e^{-\beta \cdot (R - rij)^2}$$

Ec. 10

Donde *f* es un factor de escala, *w* son las características atómicas de los átomos *i* y *j*, *rij* son las distancias interatómicas entre el átomo *i* y el átomo *j* y *A* es el número de átomos. El término exponencial contiene la distancia *rij* entre los átomos *i* y *j* y el parámetro suavizado β, que define la distribución probable de las distancias interatómicas individuales; β puede ser interpretada como un factor de

temperatura que define el movimiento de los átomos. *g(R)* es generalmente calculado como un número de puntos discretos con intervalos definidos. Algunos códigos de RDF con valores de 128 fueron propuestos, obtenidos por el ajuste del parámetro β en el rango de 100 a 200 $Å^{-2}$ y un tamaño de *R* de 0.1 a 0.2 Å.

Al incluir características de propiedades atómicas *w* de los átomos *i* y *j*, el código RDF puede ser usado en diferentes tareas ajustando los requerimientos de la información que está representando. Estas propiedades atómicas permiten la discriminación de los átomos de una molécula que no contribuyan a las características deseadas. En este caso, RDF095u, establece a *R* como 9.5 Å y la sigla *u* indica que no está ponderado por ningún término en particular. Contribuye al modelo ya que nos dice que las distancias geométricas dentro de la estructura de 9.5 Å, juegan un papel importante en el reconocimiento molecular y por lo tanto, en la estabilidad del complejo DNA- Reconocedor de surco formado. En uno de los modelos (Ecuación 8), la contribución del descriptor es negativa y por lo tanto entre mayor es este componente numéricamente, contribuye disminuyendo la Tm.

E1e. Primer componente de accesibilidad direccional, dentro de la clasificación de descriptores WHIM. Ponderado por la electronegatividad de Sanderson. Los descriptores WHIM (Weighted Holistic Invariant Molecuar descriptors) son descriptores moleculares basados en índices estadísticos calculados en las proyecciones de los átomos a lo largo de los principales ejes. Son construidos de tal forma que capturan información relevante de una molécula en 3D como el tamaño de la molécula, forma, simetría y distribución atómica con respecto a marcos de referencia invariables. El algoritmo consiste en realizar un análisis de los

principales componentes en el centro (centro: matriz molecular), dentro de las coordenadas cartesianas de una molécula, usando como ponderal la covarianza de la matriz obtenida de los diferentes esquemas ponderales de los átomos:

$$Sjk = \frac{\sum_{i=1}^{A} wi\left(qij - \bar{q}_J\right)(qik - \bar{q}_k)}{\sum_{i=1}^{A} wi}$$

Ec. 11

Donde *Sjk* es el ponderal de covarianza entre las coordenadas atómicas de *j* y *k*, *A* es el número de átomos, *wi* es el ponderal del átomo *i*, *qij* y *qik* representan las coordenadas de *j* y *k* (*j, k = x, y, z*) del átomo *i* respectivamente y *-q* el correspondiente valor de la media.

Son propuestos seis esquemas diferentes de ponderales: 1) el caso no ponderado *u* (*wi*=1, *i*=1, *n*, donde *A* es el número de átomos por cada componente), 2) masa atómica *m*, 3) El volumen *v* de fuerzas de *van der Waals,* 4) la electronegatividad atómica de *Sanderson e*, 5) La polarizabilidad atómica *p* y 6) el índice del estado electrotopológico de *Kier y Hall, S*. Todos los ponderales están en la escala con respecto al átomo de Carbono.

E1e está ponderado en las electronegatividades de Sanderson que se fundamenta en el recíproco del volumen atómico y se puede decir que parte de la interacción entre DNA-Reconocedor de surco se debe a las interacciones dipolo-dipolo, quienes están relacionada directamente con la diferencia de electronegatividad de las especies que interactúan. Este descriptor está involucrado en la ecuación 9.

D_{CL}. Como ya se mencionó, este descriptor fue obtenido a partir de la distancia de la molécula geométricamente optimizada y la topología del ADN asumiendo que se requiere una distancia específica para un reconocimiento óptimo entre ambas.

Para poder hacer una buena comparación de la serie de exploración utilizada en este estudio, se tomó una molécula reconocedora de surco menor del DNA conocida, cuyo nombre es DAPI o 4',6-diamidino-2-fenilindol (Figura 9). Esta molécula ya ha sido considerada en el estudio QSAR, con la denominación de molécula H (excluida para medir la capacidad predictiva).

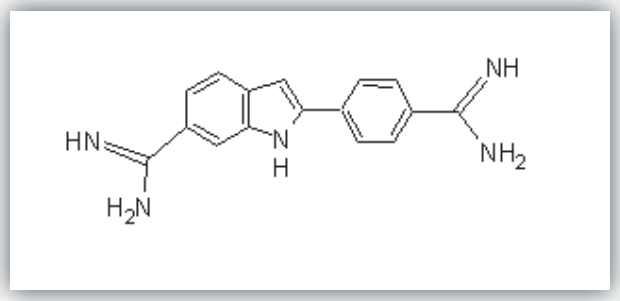


Figura 8. Estructura molecular del DAPI.

La estructura de difracción de Rayos X del DAPI unida al DNA se tomó del RCSB Protein Data Bank [42], con ID de 1D30. Es la estructura del DAPI unida a un oligonucleótido sintético (Figura10) C-G-C-G-A-A-T-T-C-G-C-G obtenida mediante difracción de Rayos X, con resolución de 2.4 Å. La estructura es casi isomorfa a la de la estructura nativa del DNA.

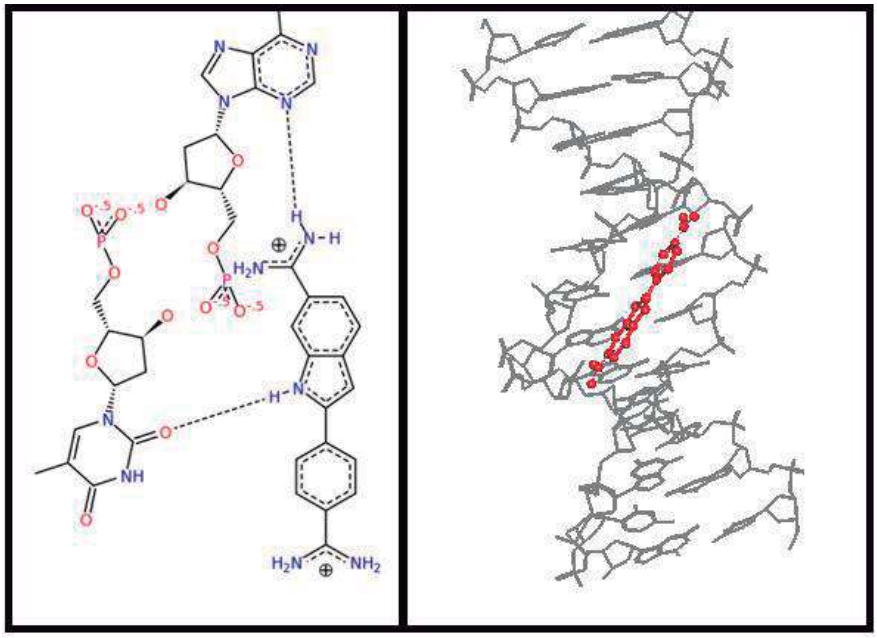


Figura 9. Interacción DNA-DAPI. Derecha: Difracción de rayos X, tomado del PDB 1D30. Izquierda: Interacción por puentes de Hidrógeno entre el DNA y DAPI.

El reconocimiento de la serie de exploración utilizada con el DNA debe ser, en teoría, similar a la del DAPI, lo cuál se comprobó al hacer un acoplamiento de todas las moléculas de la serie de datos (Figura 11), utilizando el programa ArgusLab[43] siguiendo el mismo procedimiento para todas: sobre la estructura del PDB ID 1D30, eliminando las moléculas de agua, incorporando la molécula reconocedora de surco del DNA previamente optimizada, generando el "Grupo Ligando" para el DAPI y para la molécula, y creando el sitio de unión para el DAPI para proceder con el acoplamiento donde se toman como parámetros un máximo de 150 posiciones.

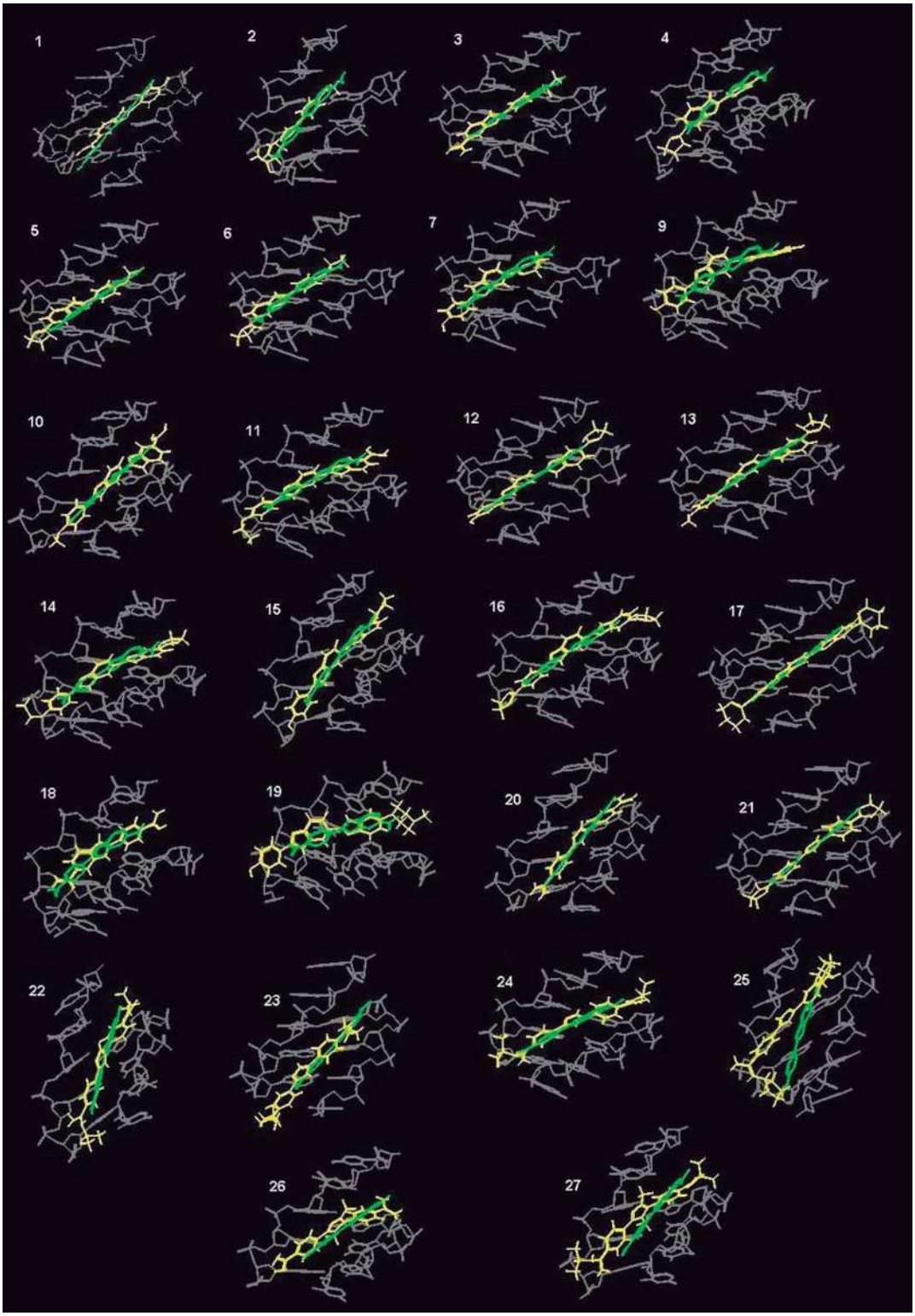


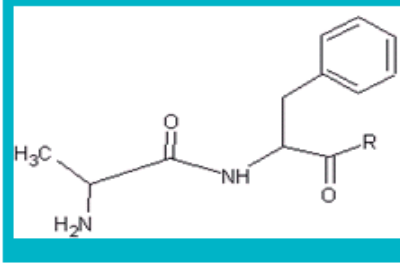
Figura 10. Acoplamiento del DNA con la serie de exploración. En amarillo el DAPI y en verde la serie de exploración.

6.2 Predicción de la actividad.

De los modelos obtenidos y de acuerdo a los descriptores moleculares que lo conforman, se puede explicar y predecir la Tm de los reconocedores de surco, aunque hay otros factores que pueden afectar el reconocimiento y la estabilidad de los reconocedores de surco y el DNA. La ecuación 7 indica que la variabilidad de Tm puede ser explicada por la helicoicidad y consideraciones geométricas, la ecuación 8 considera la helicoicidad y geometría, además de la función de distribución radial de forma negativa y finalmente, la ecuación 9 involucra electronegatividad, helicoicidad y geometría.

Los tres modelos son válidos estadísticamente aunque el que presenta mayor ventaja es el incluido en la ecuación 9, que describe de una forma más completa la interacción DNA-Reconocedor de surco, ya que contiene aspectos geométricos y electrónicos. Por esta razón, se eligió la ecuación 9 para predecir actividad de compuestos propuestos que potencialmente pueden ser considerados buenos reconocedores de surco, como derivados del dipéptido alanil-fenilalanina modificando un radical, como se observa en la tabla 6.

Tabla 6. Sustituyentes propuestos.



AP	R=H
AP1	R=$OCH_2CH_2NH_2$
AP2	R=$OCH_2CH_2N(CH_3)_2$
AP3	R=$NHCH_2CH_2NH_2$
AP4	R=$NHCH_2CH_2N(CH_3)_2$

Para realizar la predicción de actividad primero se propuso el péptido Alanil-Fenilalanina modificando el sustituyente del extremo carboxílico. Se realizó el modelado y la optimización geométrica, con los mismos criterios que la serie de exploración inicial. Posteriormente y con ayuda del programa DRAGON, se calcularon los descriptores de cada péptido propuesto. Finalmente se aplica la ecuación 9, obteniendo el log (ΔTm) (Tabla 7).

Tabla 7. Predicción de actividad del péptido propuesto.

	G(N..N)	E1e	D_{CL}	log (ΔTm)
AP1	18.19100	0.41800	0.063	0.791
AP2	18.18400	0.48700	0.070	0.906
AP3	28.58400	0.43100	-0.099	0.809
AP4	29.21500	0.49000	-0.017	0.979

Con fines comparativos, se incluye la ecuación 12 del modelo 3 de la tabla de resultados (Tabla 3), para realizar la predicción de actividad de los péptidos propuestos.

$$\log(\Delta Tm) = 0.437ATS2m - 0.331MATS8e + 0.448\,MATS8p - 0.567GATS1p - 3.312$$

n= 27 R^2 = 0.9551 F=107.63 s=0.026 Q^2= 92.66 **Ec. 12**

La actividad calculada con la ecuación 9 y con la ecuación 12 se muestra en la tabla 8, así como la diferencia que hay entre los valores calculados con cada una de las ecuaciones.

Tabla 8. Comparación entre actividad experimental y calculada con la ecuación 9 y 12.

		Ecuación 9	Ecuación 12	
Mol.	log (ΔTm) experimental	log (ΔTm) calculado	log (ΔTm) calculado	Diferencia
1	0.845	0.783	0.796	0.013
2	0.792	0.814	0.789	0.025
3	0.69	0.743	0.757	0.014
4	0.748	0.789	0.800	0.011
5	0.875	0.768	0.881	0.113
6	0.845	0.77	0.784	0.014
7	0.845	0.808	0.856	0.048
8	1.233	1.405	1.075	0.330
9	1.362	1.4	1.375	0.025
10	1.362	1.267	1.421	0.154
11	1.362	1.369	1.291	0.078
12	1.378	1.394	1.341	0.053
13	1.387	1.408	1.397	0.011
14	1.412	1.323	1.288	0.035
15	1.405	1.266	1.388	0.122
16	1.479	1.482	1.530	0.048
17	1.487	1.477	1.591	0.114
18	1.233	1.083	1.260	0.177
19	1.253	1.356	1.261	0.095
20	0.959	0.978	0.948	0.030
21	0.857	0.998	0.868	0.130
22	1.037	1.048	0.931	0.117
23	1.037	1.024	1.049	0.025
24	0.968	0.913	0.931	0.018
25	1.113	1.157	1.081	0.076
26	0.813	0.862	0.801	0.061
27	0.732	0.911	0.837	0.074

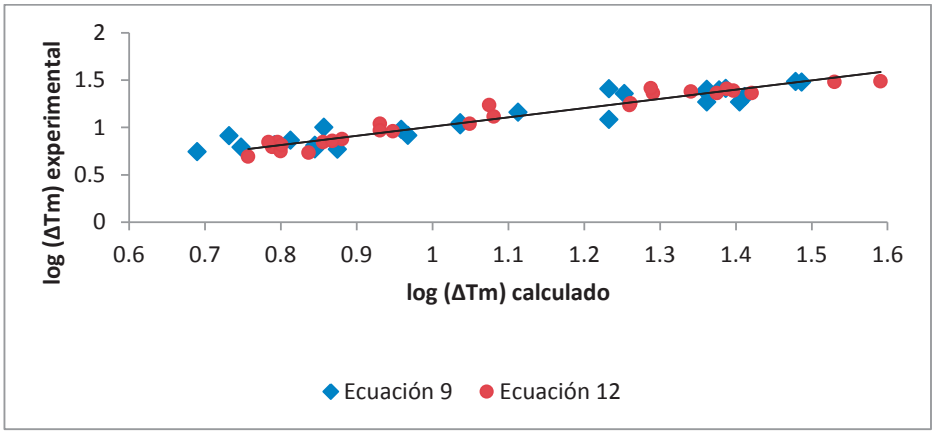


Figura 11. Dispersión entre la actividad experimental y calculada de la ecuación 9 y la ecuación 12.

Utilizando la ecuación 12 se calcula la actividad de los péptidos propuestos, cuyos resultados se muestran el la tabla 9.

Tabla 9. Actividad calculada del péptido propuesto con la ecuación 12.

Péptido	Ecuación 9	Ecuación 12
AP1	0.791	0.381
AP2	0.906	0.373
AP3	0.809	0.414
AP4	0.979	0.401

Aunque se observa una diferencia considerable entre ambas predicciones, resulta indispensable llevar a cabo una validación externa que permita establecer el mejor modelo. Sin embargo cabe resaltar que el estudio QSAR desarrollado en este trabajo tiene como finalidad utilizarse como modelo explicativo y no predictivo.

El modelo 3 que da origen a la ecuación 12, está compuesto por descriptores de la familia 2D autocorrelaciones, que explican de forma general como ciertas propiedades son distribuidad a lo largo de la estructura topológica. Los vectores de autocorrelación fueron calculados en un espacio que va de 1 a 8. La tabla 10 resume el significado de cada descriptor involucrado en la ecuación 12.

Tabla 10. Interpretación de los descriptores de la ecuación 12.

Descriptor	Interpretación
ATS2m	Autocorrelación Broto-Moreau de una estructura topológica. Ponderado por masa atómica.
MATS8e	Autocorrelación Moran. Ponderado por electronegatividad atómica de Sanderson.
MATS8p	Autocorrelación Moran. Ponderado por polarizabilidad atómica.
GATS1p	Autocorrelación Geary. Ponderado por polarizabilidad atómica.

Tomando como base la Tabla 7, la actividad predicha por el modelo QSAR permite recomendar, por conveniencia, la síntesis de los derivados propuestos aunque se consideró importante contar con más elementos para la elección adecuada. Es por esta razón que se llevó a cabo un estudio de anclaje molecular (Docking) de los compuestos propuestos. Los resultados se describen a continuación.

6.3 Acoplamiento del péptido en el DNA.

Se realizó el acoplamiento entre el péptido con sus diferentes sustituyentes mediante el uso de anclaje molecular "Docking", con los mismos criterios utilizados en el apartado anterior. El acoplamiento del péptido en sus diferentes combinaciones, del AP1 al AP4 se observan en las figuras 12, 13, 14 y 15 respectivamente. En ellas se observa la superposición del péptido con el DAPI y la conformación que adoptan para el acoplamiento óptimo en el surco menor del DNA.

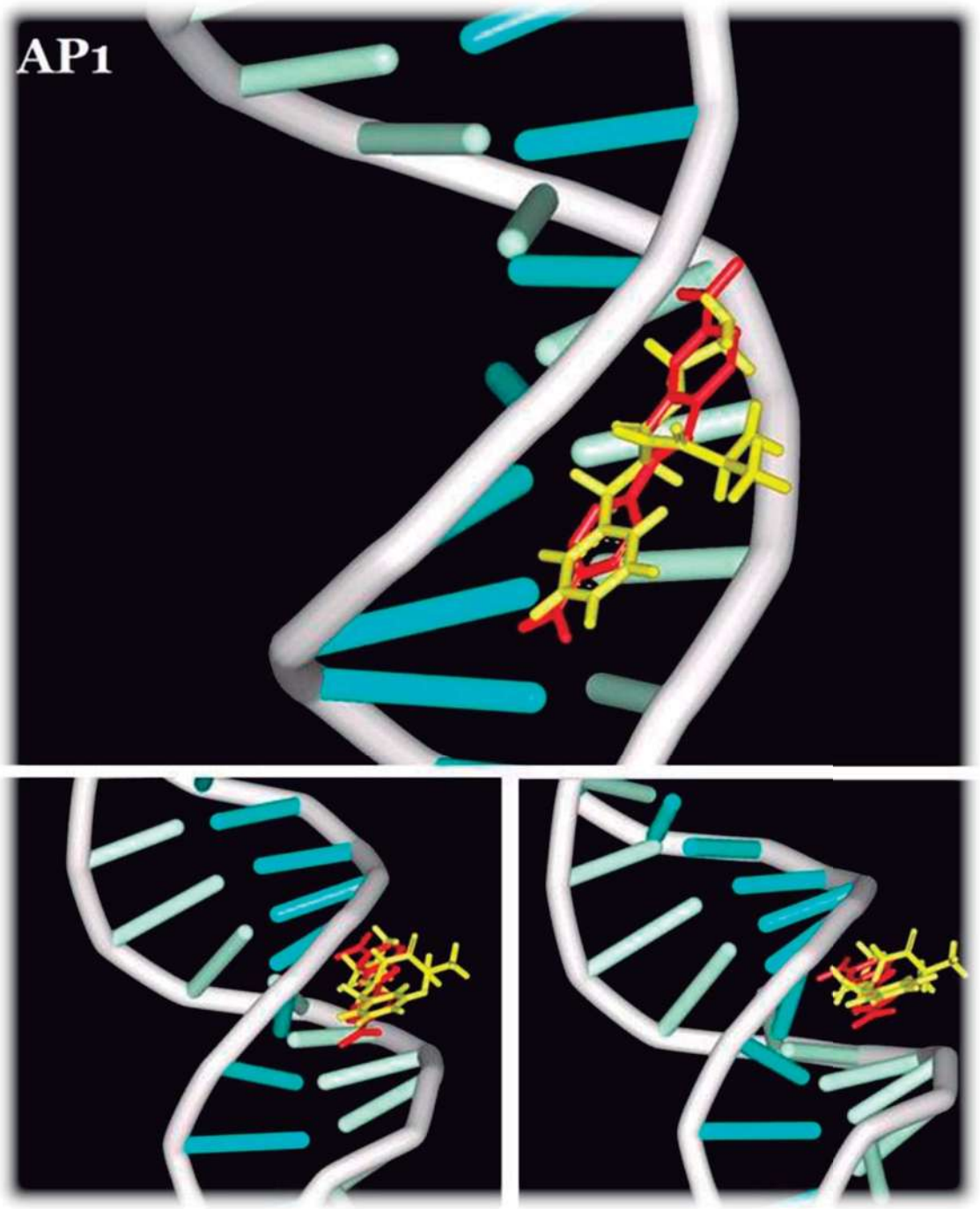


Figura 12. Acoplamiento del péptido AP1 al DNA y la superposición con el DAPI como referencia. En amarillo el péptido y en rojo el DAPI.

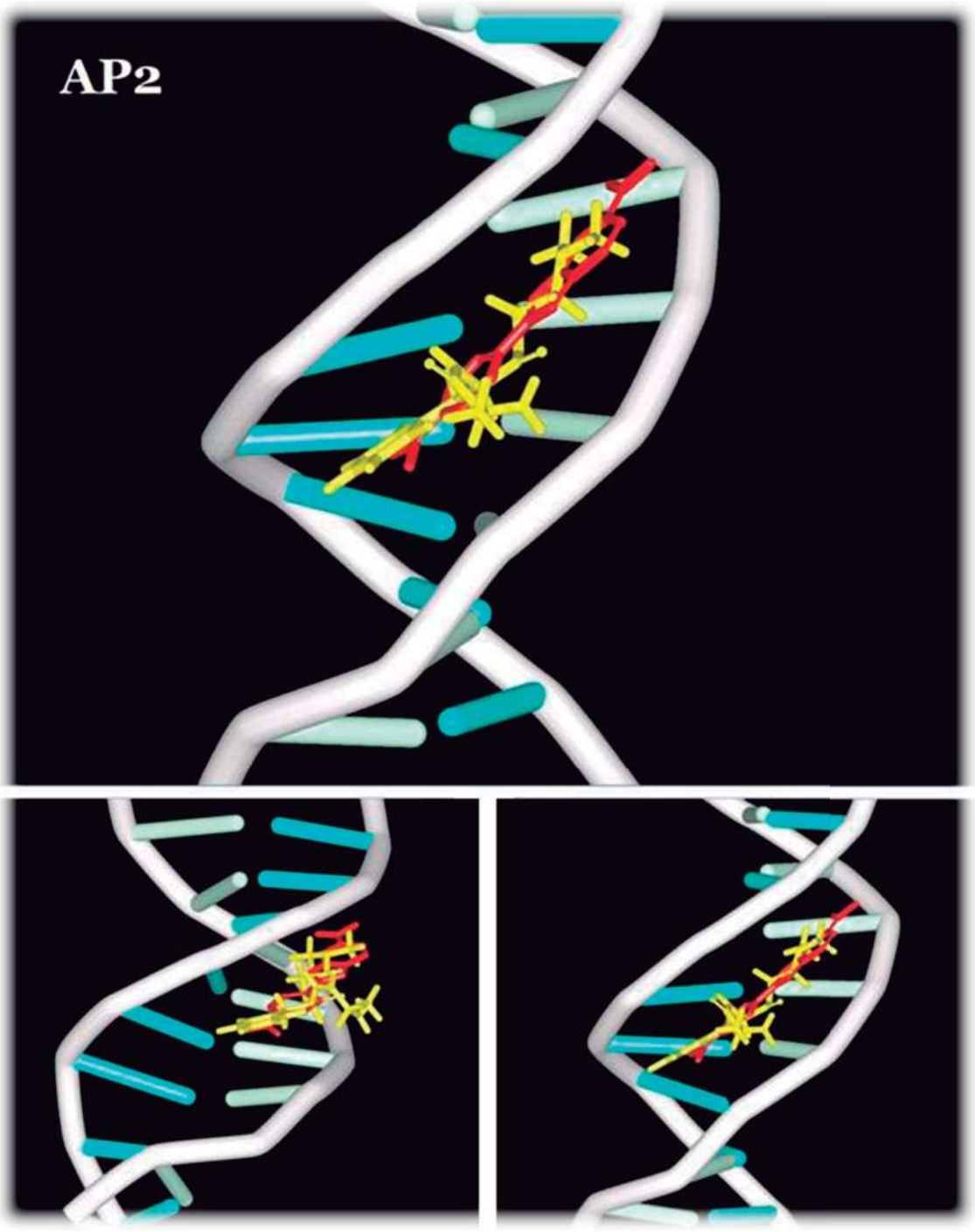


Figura 13. Acoplamiento del péptido AP2 al DNA y la superposición con el DAPI como referencia. En amarillo el péptido y en rojo el DAPI.

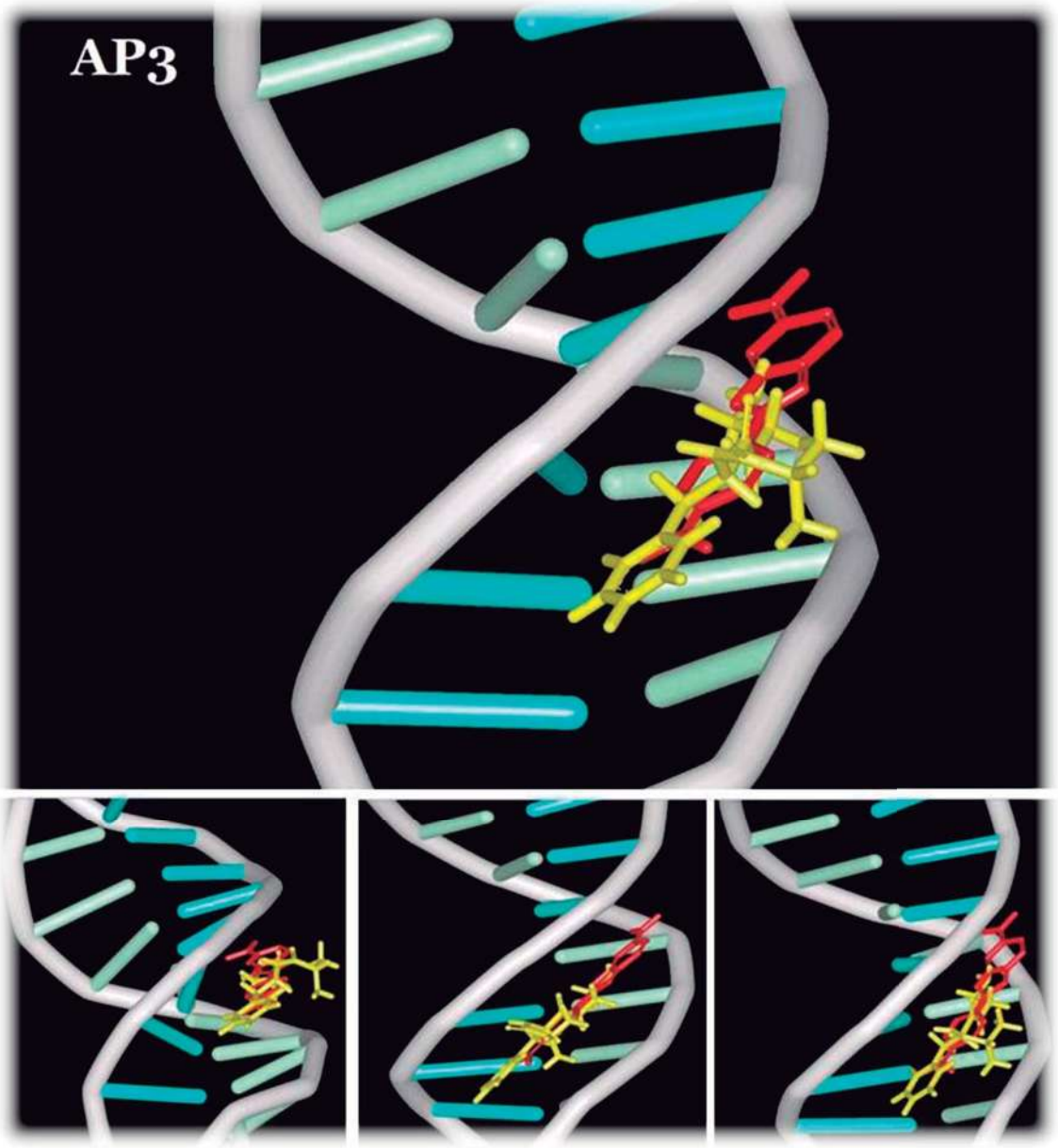


Figura 14. Acoplamiento del péptido AP3 al DNA y la superposición con el DAPI como referencia. En amarillo el péptido y en rojo el DAPI.

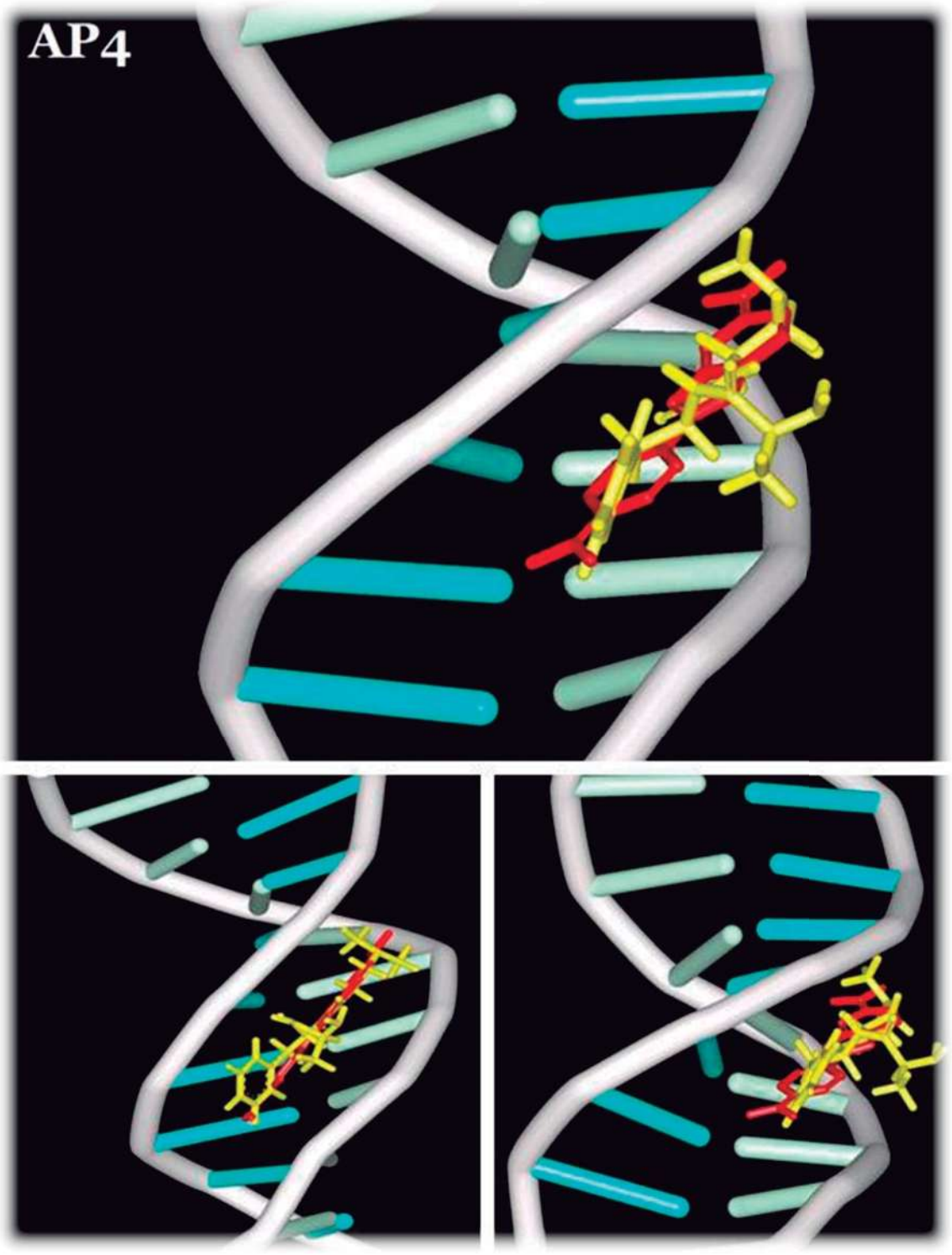


Figura 15. Acoplamiento del péptido AP4 al DNA y la superposición con el DAPI como referencia. En amarillo el péptido y en rojo el DAPI.

Se observó que la interacción del DNA con el péptido se encuentra en zonas ricas en bases A-T (figura 16) lo que confirma que la interacción sigue los principios de un reconocedor de surco típico. El sitio de unión o "Binding Site" es donde se encuentra el DAPI y como consecuencia, será el sitio de acoplamiento esperado para el péptido. La secuencia que sigue el sitio de unión no es complementaria entre una cadena y otra; la secuencia es de 5'-^{1} A^{6}T^{7}T^{8}C^{9} ^{12}-3' y la secuencia complementaria es de 3'-^{24} C^{21}T^{20}T^{19}A^{18}A^{17} ^{13}-5'.

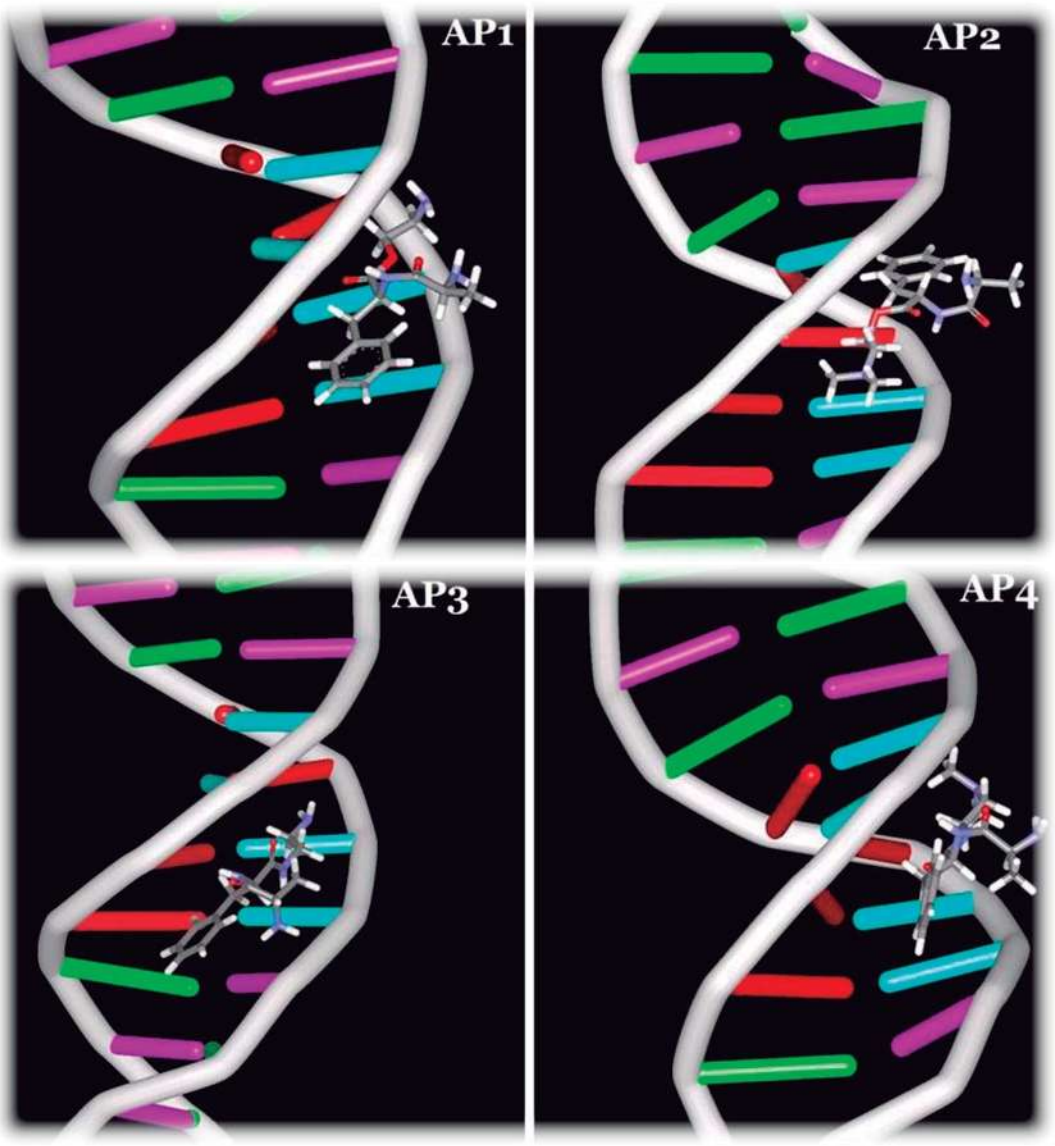


Figura 16. Regiones de acoplamiento ricas en A – T.

Cuando se realiza el acoplamiento, la energía del complejo DNA-Reconocedor de surco pasa por diversas "poses" para buscar la energía óptima y al llegar a ésta, queda unido el ligando y se forma el complejo. El número de poses realizadas y la energía utilizada se resumen en la tabla 8.

Tabla 11. E_c óptima de formación del complejo DNA-Ligando.

Ligando	No. Poses	E_c óptima (kcal/mol)	E_c alta (kcal/mol)	log (ΔTm)
AP1	53	-3.9	-3.63	0.791
AP2	64	-3.93	-3.55	0.906
AP3	47	-4.07	-3.79	0.809
AP4	84	-4.06	-3.51	0.979
DAPI*		-5.08		1.233

*Valores de DAPI tomados de la literatura[36]

Las interacciones que se esperan entre el ligando (que en este caso es el péptido) y el DNA son puentes de Hidrógeno. En la figura 17 se observa la interacción por puentes de Hidrógeno que presenta el péptido AP3 con la base nitrogenada Timina. Además, en la figura 18 se observa la posición dentro del DNA y la conformación que adopta en el surco menor.

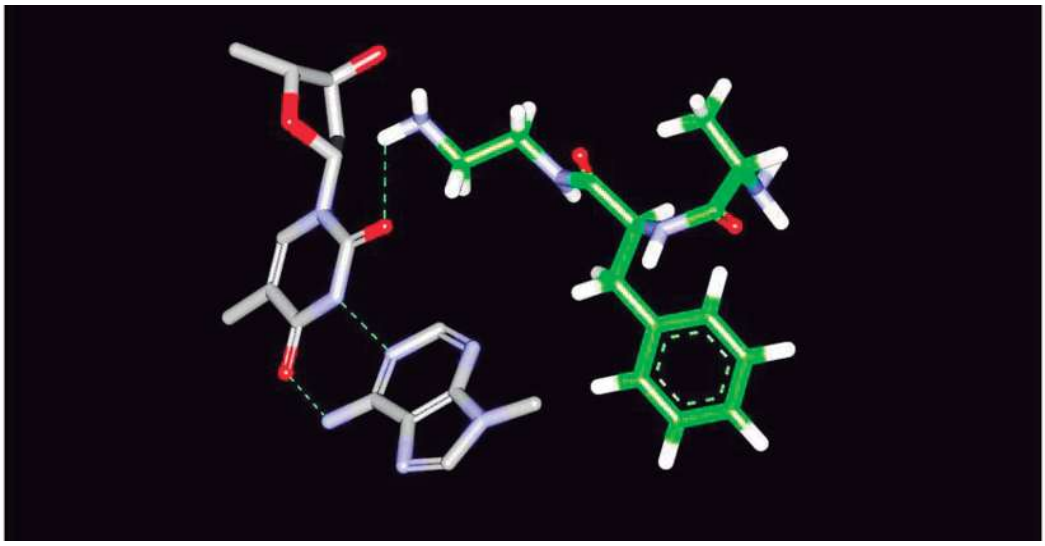


Figura 17. Puentes de Hidrógeno entre AP3 y Timina.

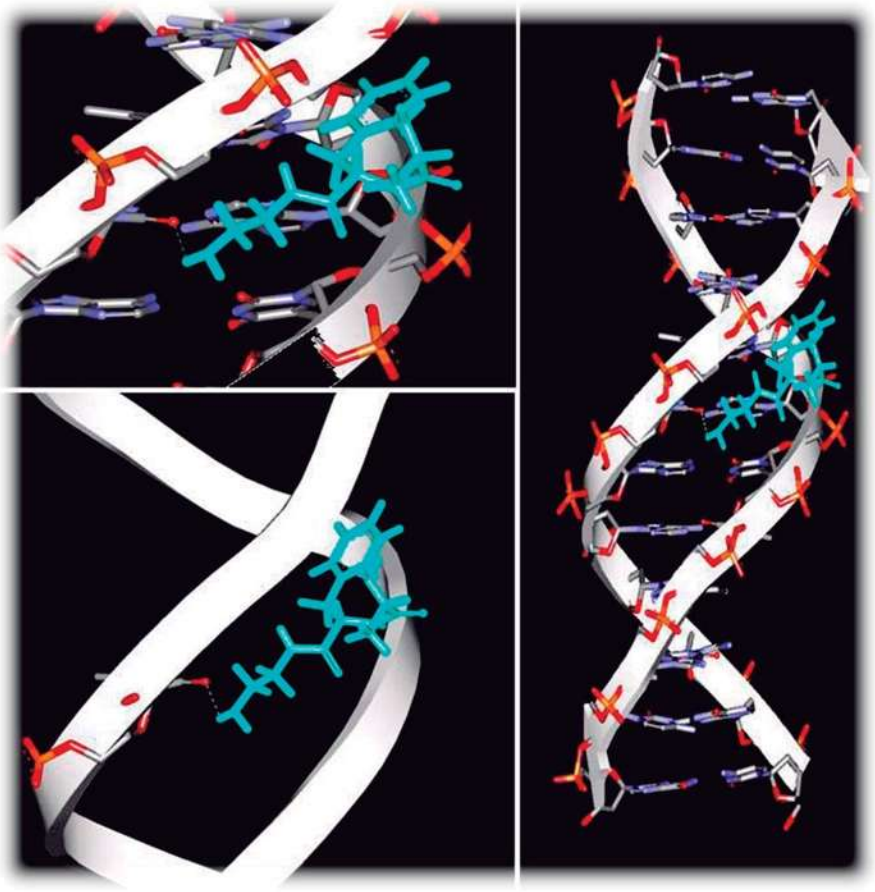


Figura 18. Interacción por puente de Hidrógeno de AP3 y el DNA

Al comparar con el DAPI, el péptido que más se acerca a los valores óptimos para el reconocimiento en el DNA es el AP4, ya que tiene una buena E_c y además, es el que presenta mayor log (ΔTm). Se observa una relación proporcional entre la energía óptima y el tamaño de la molécula; entre más grande el reconocedor de surco, más estabilidad en el sistema. Esto se explica gracias a que un reconocedor de surco menor toma una forma de media luna para poder tener interacción mediante puentes de Hidrógeno y el tamaño toma un papel importante para adoptar dicha forma geométrica. También se observa que hay una preferencia por el N más que por el O de los radicales.

7. CONCLUSIONES

El estudio QSAR mediante algoritmos genéticos aporta modelos con información topológica, geométrica y electrónica que permiten predecir actividad de reconocedores de surco distintos a los utilizados en este estudio.

El descriptor D_{CL} se integra a los modelos estadísticamente significativos para predecir actividad teórica, como un descriptor con información topológica de helicoicidad.

Se puede demostrar la utilidad del descriptor D_{CL} como una característica importante para la formación óptima del complejo DNA- Ligando mediante el acoplamiento de reconocedores de surco teóricamente activos como reconocedores de surco.

La energía de acomplejamiento E_c se puede utilizar como un marcador para llevar a cabo un buen acoplamiento del reconocedor de surco y el DNA, comprobando que se requiere una distancia específica para que el reconocimiento entre macromolécula y el ligando pueda llevarse a cabo.

Se recomienda la síntesis del derivado peptídico AP4 para ser probado como reconocedor de surco.

8. BIBLIOGRAFÍA

[1]Cuevas G., Cortés F. 2003. **Introducción a la química computacional**. 1a. ed. *Fondo de Cultura Económica*. México, D.F. pp. 23-24.

[2]Alemán C., Muñoz-Guerra S. 2003. **Aplicaciones de los métodos computacionales al estudio de la estructura y propiedades de polímeros**. *Polímeros: Ciência e Tecnologia* 4 (13): 250 - 264.

[3]Stewart J. J. P. 1989. **Optimization of parameters for semiempirical methods I. Method.** *Journal of Computational Chemistry* 2 (10): 209 - 220.

[4]Dewar M. J. S., Zoebisch E. G., Hearly E. F., Stewart J. J. P. 1985. **The development and use of quantum mechanical molecular models. 76. AMI: a new general purpose quantum mechanical molecular model.** *Journal of the American Chemical Society*. 107: 3902 - 3909.

[5]Nicolás-Vázquez M. I., Marín C. E., Castro M. F. M., Miranda R. R. 2006. **Algunos aspectos básicos de la química computacional.** 1a. ed. *Universidad Nacional Autónoma de México*. Cuatitlán Izcalli, Estado de México. pp. 37 - 38.

[6]Gago Badenas F. 1994. **Métodos computacionales de modelado molecular y diseño de fármacos.** *Monografía I. Diseño de medicamentos*. Real Academia Nacional de Farmacia. Universidad de Alcalá de Henares, Madrid. pp. 253 – 311.

[7]González J. E., Poltev V. I. 2002. **La simulación computacional de procesos genéticos a nivel molecular.** *Elementos* 47. pp. 31 - 35.

[8]Tereshko V., Minasov G., Egli M. 1999. **The Dickerson-Drew B-DNA Dodecamer revisited atatomic resolution**. *American Chemical Society* 2 (121): 470 - 471.

[9]Ondarza R. N. 1994. **Biología molecular: antes y después de la doble hélice**. 1a. ed. Siglo XXI editors, S.A. de C.V. México, D.F. pp. 54 - 57.

[10]Kumar S.H., Chourasia M., Kumar D., Narahari S. G. 2011. **Comparasion of computational methods to model DNA minor groove binders.** *Journal Chemical Information and Modeling* 51: 558 - 571.

[11]Schalley C. A., Springer A. 2009. **Mass spectrometry and gas-phase chemistry of non-covalent complexes.** John Wiley & Sons, Inc. U.S.A. pp. 477 - 480

[12]Rueda M., Luque F. J., Orozco M. 2005. **Nature of minor-groove binders-DNA complexes in the gas phase.** *American Chemical Society* 33 (127): 11690 - 11698.

[13]Degtyareva N. N., Wallace B. D., Bryant A. R., Loo K. M., Petty J. T. 2007. **Hydration changes accompanying the binding of minos groove ligands with DNA**. *Biophysical Journal* 92: 959 - 965.

[14]Bürli R. W., Ge Y., White S., Baird E. E., Touami S. M., Taylor M., Kaizerman A., Moser H. E. 2002. **DNA binding ligands with excellent antibiotic potency against drug-resistant gram-positive bacteria**. *Bioorganic & Medicinal Chemistry Letters* 12: 2591 - 2594.

[15]Khan G. S., Shah A., Rehman Z., Barker D. 2012. **Chemistry of DNA minor groove binding agents**. *Journal of Photochemistry and Photobiology B: Biology 3 (115):105 – 108.*

[16]Boykin D. W. 2002. **Antimicrobial Activity of the DNA Minor Groove Binders Furamidine and Analogs**. *Journal of the Brazilian Chemical Society* 6 (12): 763 - 771.

[17]Maloy S. R., Cronan J. E., Freifelder D. 1994. **Microbial Genetics**. 2nd. ed. Courier Corporation. U.S.A.. pp. 33 - 34.

[18]Lewin B. 1996. **Genes.** Vol. 1. Editorial Reverté, S.A. Barcelona, España. pp. 98 - 100.

[19]Le Novère N. 2001. **MELTING computing the melting temperature of nucleic acid dúplex**. *Oxford University Press* 12 (17): 1226 - 1227.

[20]Campos P., Sanmartí, Torres, Mingo, Fernández, Boixaderas, De la Rubia, Rodríguez, Pintó, Gullón. 2002. **Biología 2**. Editorial Limusa S.A de C.V. México, D. F. pp. 25.

[21]Solari A. J. 2004. **Genética humana: Fundamentos y aplicaciones en medicina**. 3a. ed. Editorial Médica Panamericana S.A. Buenos Aires, Argentina. pp. 143

[22]Devlin T. M. 2004. **Bioquímica: Libro de texto con aplicaciones clínicas**. Editorial Reverté S.A., Barcelona, España. pp. 186.

[23]Zavala-Franco A., González C. J. B., Chacón G. L. 2010. **Estudio QSAR por algoritmos genéticos de reconocedores de surco del DNA.** *Biológicas* 12 (2): 108 - 115.

[24]Cerralaza M., Annicchiarico W. 1996. **Algoritmos de Optimización Estructural Basados en Simulación Genética.** Universidad Central de Venezuela. Caracas, Venezuela. pp. 45 - 47.

[25]Quintero R. H. F., Calle T. G., Díaz A. A. 2004. **Síntesis de generación de trayectoria y de movimiento para múltiples posiciones en mecanismos, utilizando algoritmos genéticos**. *Scientia Et Technica* 25 (10): 131 - 136.

[26]Melián B. B., Moreno P. J. A., Moreno V. J. M. 2009. **Algoritmos genéticos: una visión práctica.** *Revista Didáctica de las Matemáticas* 71: 29 - 47.

[27]Puzyn T., Leszczynski J., Cronin T. D. **Recent Advances in QSAR Studies: Methods and Applications**. Editorial Springer. New York, U.S.A. pp. 1 - 45.

[28]Prado-Prado F. J., Martínez de la Vega O., Uriarte E., Ubeira FM., Chou K.C., González-Díaz H. 2009. **Unified QSAR approach to antimicrobials. 4. Multi-target QSAR modelling and comparative multi-distance study of the giant components of antiviral drug–drug complex networks**. *Bioorganic & Medicinal Chemistry* 17 (2): 569–575

[29]García Ramos J.C., Bravo M.E., Ortiz Frade L.A., Ruiz Azuara L. **Estudio QSAR de compuestos de coordinación de cobre de tipo [Cu(N-N)(glicinato)] NO_3.** 2° Congreso Nacional de Química Médica.

[30]Goodarzi M., da Cunha EF., Freitas MP, Ramalho TC. 2010. **QSAR and docking studies of novel antileishmanial diaryl sulfides and sulfonamides**. *European Journal of Medicinal Chemistry* 45 (11): 4879 – 4889.

[31]Vicente E., Duchowicz PR., Benítez D., Castro EA., Cerecetto H., González M., Monge A. 2010. **Anti-T. cruzi activities and QSAR studies of 3-arylquinoxaline-2-carbonitrile di-N-oxides.** *Bioorganic & Medicinal Chemistry Letters* 20 (16): 4831–4835.

[32]Ho Cho D., Kwang Lee S., Tae Kim B., Tai No K.. 2001. **Quantitative Structure-Activity Relationship (QSAR) Study of New Fluorovinyloxyacetamides.** *Bulletin of the Korean Chemical* Society 4 (22): 388 - 394.

[33]Escalona J.C., Carrasco R., Padrón J. A. 2008. **Introducción al diseño de fármacos. Folleto para la docencia de la asignatura de Farmacia**. Universidad de Oriente. La Habana, Cuba.

[34]Pastor M., Alvarez-Builla. **Técnicas QSAR en diseño de fármacos.** Departamento de Química Orgánica, Universidad de Alcalá. Alcalá de Henares. Madrid. pp. 69-98.

[35]Chacón-García L., Martínez R. 2001. **Cytotoxic activity and QSAR of N,N'-diarylalkanediamides.** *European Journal of Medicinal Chemistry* 36: 731–736.

[36]De Oliveira A. M., Custódio F. B., Donicci C. L., Montanari C. A. 2003. **QSAR and molecular modelling studies on B-DNA recognition of minor groove binders.** *European Journal of Medicinal Chemistry* 38: 141 - 155.

[37]**HyperChem** 6.03 for Windows; HyperCube Inc., 2000.

[38]Press W, H., et. al., 1968. **Numerical Recipes: The Art of Scientific Computing**. *Cambridge University Press*. Chapter 10.

[39]Todeschini R., Consonni V., Mauri A. and Pavan M. 2005. **DRAGON—Software for the calculation of molecular descriptors.** Ver. 5.3, Talete srl, Milano, Italy.

[40]Todeschini R., Ballabio D., Consonni V., Mauri A. and Pavan M. 2004. **Mobydigs Computer Software**, 1.0; TALETE srl: Milano.

[41]Software **Microsoft Office Excel** 2010. 2010. Para Windows. Microsoft Corporation.

[42]**RCSB Protein Data Bank** URL:http://www.rcsb.org/pdb/home/home.do

[43]**ArgusLab 4.0.1**. 1997 - 2004. Mark Thompson and Planaria Software LLC.

APÉNDICE A

Electronegatividad atómica de Sanderson.

R. T. Sanderson propone que la electronegatividad de un átomo es una medida de la "densidad" de su nube electrónica, comparada con la de un átomo inerte hipotético con el mismo número de electrones, ya que si un átomo es muy electronegativo, es decir, atrae fuertemente a otros electrones, los suyos propios los mantendrá muy próximos entre sí. Sanderson define la densidad electrónica media (DE) mediante la relación:

$$DE = \frac{3Z}{4\pi r^2}$$

Donde *r* es el radio covalente en amstrongs.

Electronegatividad de Kier y Hall.

Índices de conectividad de valencia. Describe topología.

$$\mathrm{m_{X^v}} = \sum_{i=1}^{N_s} \prod_{k=1}^{m+1} \left(\frac{1}{\delta_k^v}\right)^{1/2}$$

$\delta_k^v = \frac{(Z_k^v - H_k)}{(Z_k - Z_k^v - 1)}$ = valencia de conectividad para el átomo *k-th* en el gráfico molecular.

Z_k -Número total de electrones en el átomo *k-th*.

Z_k^v -Número de electrones de valencia en el átomo *k-th.*

H_k -Número de átomos de hidrógeno conectados directamente al *k-th*

m = 0 –índice de conectividad de la valencia atómica.

m = 1 –Índice de conectividad de un enlace de valencia.

m = 2 –índice de conectividad de dos fragmentos de enlace de valencia.

m = 3 –índice de conectividad de tres fragmentos contiguos de enlace de valencia.

Autocorrelación Broto-Moreau

Es la autocorrelación espacial más conocida, definida en una gráfica molecular *G* como:

$$ATS_k = \frac{1}{2} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} w_i \cdot w_j \cdot \delta(d_{ij}; k) = \frac{1}{2} \cdot (w^T \cdot kB \cdot w)$$

Donde *w* es una propiedad atómica, *A* es el número de átomos en la molécula, *k* es el espacio, y *dij* es la distancia topológica entre los átomos *i* y *j*, $\delta(d_{ij}; k)$ es una función delta Kronecker igual a 1 si *dij=k*, 0 en cualquier otro caso y *kB* es el orden de *kth.*

Autocorrelación Moran.

Índice general de autocorrelación espacial que, si se aplica a una gráfica molecular, puede ser definida como:

$$I(d) = \frac{\frac{1}{\Delta} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} \delta ij.\,(wi - \overline{w}) \cdot (wj - \overline{w})}{\frac{1}{A} \cdot \sum_{i=1}^{A} (wi - \overline{w})^2}$$

Donde *wi* es una propiedad atómica, $\overline{w}$ es el valor promedio en la molécula, *A* es el número atómico, *d* es la distancia topológica considerada, δij es un valor delta Kronecker y Δ es la suma de los deltas Kronecker (número de pares de vértices en una distancia igual a *d*).

El coeficiente Moran frecuentemente toma un valor del intervalo [-1,+1]. La autocorrelación positiva corresponde a valores positivos, mientras que la autocorrelación negativa produce valores negativos.

Autocorrelación Geary

Índice general de la autocorrelación espacial que, si se aplica a una gráfica molecular, puede ser definida como:

$$c(d) = \frac{\frac{1}{2\Delta} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} \delta ij.\,(wi - wj)^2}{\frac{1}{(A-1)} \cdot \sum_{i=1}^{A} (wi - \overline{w})^2}$$

Donde *wi* es una propiedad atómica, $\overline{w}$ es el valor promedio en la molécula, *A* es el número atómico, *d* es la distancia topológica considerada, δij es un valor delta Kronecker y Δ es la suma de los deltas Kronecker (número de pares de vértices en una distancia igual a *d*).

El coeficiente Geary es una función de distintos tipos de distancias de cero a infinito. Una fuerte autocorrelación produce altos valores de este índice, por otra parte, autocorrelaciones positivas se traducen en valores entre 0 a 1, mientras que autocorrelaciones negativas producen valores grandes donde 1, por lo tanto, hace referencia a "no correlación" cuando *c* es igual a 1.

Estudio QSAR por algoritmos genéticos de reconocedores de surco del DNA

Anai Zavala Franco, Janett Betzabe González Campos y Luis Chacón García*

Laboratorio de Diseño Molecular, Instituto de Investigaciones Químico Biológicas, Edificio B-1, Ciudad Universitaria, Morelia, Michoacán, México. CP 58066

Resumen

Se describe un análisis de Relación Estructura Actividad Cuantitativa (QSAR) por Inteligencia Artificial mediante el uso de algoritmos genéticos de una serie de reconocedores de surco del DNA. Se encuentra una correlación entre las características estructurales y geométricas de reconocimiento intermolecular y la Tm del complejo ligando-DNA.

Palabras Clave: *QSAR, Reconocedores de surco del DNA, descriptores moleculares.*

Abstract

A quantitative structure activity relationship (QSAR) by the use of genetic algoritms of a series of DNA-groove binders is described. It was found a correlation between Tm of the DNA-recognizing molecule and its structural and geometric properties.

Keywords: *QSAR, DNA-groove binders, molecular descriptors.*

Introducción

Un algoritmo es por definición una serie de pasos organizados para resolver un problema específico. Durante la última década se ha incrementado el interés en el uso de algoritmos en las ciencias biológicas gracias a los avances en la computación. Una de las técnica de computación evolucionaria, son los algoritmos genéticos [Estévez, 1997] que toman como inspiración la evolución genética y sus bases moleculares.

Los Algoritmos genéticos (AGs) son métodos adaptativos que pueden usarse para resolver problemas de búsqueda y optimización. Están basados en el proceso mediante el cual los organismos vivos, a lo largo de las generaciones, evolucionan de acuerdo a los principios de la selección natural y la supervivencia del "más fuerte", postulados por Darwin. Los Algoritmos Genéticos, un área de las ciencias computacionales y las matemáticas imitan este proceso, lo que facilita buscar una solución a problemas reales. La evolución de dichas soluciones hacia valores óptimos del problema depende de una buena codificación de las mismas.

En la naturaleza los individuos de una población compiten entre sí en la búsqueda de recursos tales como comida, agua, refugio, incluso un compañero. Aquellos individuos que tienen más éxito en sobrevivir y en atraer compañeros tienen mayor probabilidad de generar un gran número de descendientes. Por el contrario, individuos poco dotados producirán un menor número de descendientes. Es posible pensar que los genes de los individuos mejor adaptados se propagarán en sucesivas generaciones hacia un número de individuos creciente. La combinación de buenas características provenientes de diferentes ancestros, puede a veces producir descendientes cuya adaptación es mucho mayor que la de cualquiera de sus ancestros.

Los AGs utilizan este razonamiento, de tal manera que se parte de una población inicial, siendo sus individuos una posible solución a un determinado problema. A cada individuo se le asigna un valor (en la naturaleza, es el grado de efectividad de un organismo para competir por determinados recursos). Mientras más adaptado esté el individuo al problema, mayor será la probabilidad de que sea seleccionado para reproducirse, cruzando su material genético con otro seleccionado de la misma forma.

De esta manera se produce una nueva población de posibles soluciones, que contiene mejores características para una posible solución final satisfactoria. Así, a lo largo de las generaciones, las buenas características se propagan a través de la población. [Larrañaga *et al.*)]

Una de las disciplinas que han aprovechado la inteligencia artificial y particularmente los AG´s es la Química Farmacéutica dentro del diseño de fármacos a través de los métodos de relación estructura actividad cuantitativa o QSAR (por sus siglas en inglés Quantitative Structure-Activity Relationships)[Pastor, *et al.*] que han agrupado todas las técnicas intentando establecer modelos empíricos de comportamiento, sobre familias de compuestos biológicamente activos, como por ejemplo un grupos de no-nucleósidos inhibidores de VIH-1 [Hou T. J., *et al.*, 1999], para poder obtener óptimos de actividad, a partir datos de comportamiento de un número limitado de compuestos. En resumen, un análisis QSAR busca una relación matemática entre las características fisicoquímicas de una serie de compuestos activos y su actividad biológica cuantificada, lo que permite explicar la actividad de dichos compuestos y más aún predecir la actividad de compuestos no sintetizados o aun no probados biológicamente. Así, es posible citar estudios realizados de QSAR en productos naturales (p. ejemplo flavonoides) [Stefanic-Petek, *et al.*, 2002], agentes antibacterianos [Koga, *et al.*, 1980], compuestos antifúngicos [Yalcin, *et al.*, 2000], estudios de toxicidad [Turabekova, *et al.*, 2004], compuestos citotóxicos [Suh, *et al.*, 2002], entre otros.

Un grupo de compuestos con actividad farmacológica de interés son los que interactúan con el DNA, dado su uso potencial como agentes citotóxicos o antibacterianos. [Martínez y Chacón-García, 2005] Por su mecanismo de acción molecular, estos se pueden clasificar como agentes que interactúan

✉***Autor de correspondencia:** Luis Chacón García, e-mail: lchacon@umich.mx

covalentemente con el material genético y aquellos que lo hacen de manera reversible mediante interacciones intermoleculares débiles tales como puentes de hidrógeno, fuerzas de van der Waals o interacciones electrostáticas. Estos últimos son de gran importancia pues, cuando se trata de fármacos, su reversibilidad en la interacción permite o facilita la destoxificación de la célula y por consiguiente del organismo. A su vez, se subdividen en intercaladores y en reconocedores de surco (RS) dependiendo del tipo de interacción con el DNA. **Figura 1**. [Martínez y Chacón-García, 2005]

Los RS, al ser compuestos que reconocen al material genético de manera reversible, están sujetos a los principios básicos de la termodinámica, de tal manera que la energía requerida para su interacción se define por la Ecuación 1:

$$\Delta G = \Delta H - T\Delta S \qquad \textbf{Ecuación 1}$$

En donde *G* representa la energía libre de Gibbs, *H* la entalpía, *S* la entropía y T la temperatura del sistema que se considera constante.

Dada la ecuación 1, se deduce que el proceso será termodinámicamente favorecido (ΔG negativo) cuando el cambio en la entalpía se ve favorecido, por lo que a su vez un aumento en las interacciones intermoleculares estabilizantes (Δ*H* negativo) con el DNA producirá un proceso espontáneo y deseado para su óptimo reconocimiento intermolecular. Dentro de estas interacciones se encuentran los puentes de hidrógeno que son los que gobiernan la interacción de los RS con las bases púricas y pirimídicas.

Los RS pueden ser reconocedores, en el B-DNA, del surco mayor o del menor. [Neidle, 2001] El surco mayor es principalmente reconocido por proteínas y aunque el surco menor es considerado menos específico es el responsable del reconocimiento de moléculas orgánicas pequeñas. [Bewley, *et al.*, 1998] Estas últimas adoptan, en su sitio de acción molecular, una conformación helicodal.

En un estudio previo, se describió el efecto de la isohelicoicidad, en función de sus interacciones por puentes de hidrógeno, de una serie de N,N'-diarilalcanodiamidas con longitud variable y con actividad citotóxica moderada. Se propuso un descriptor molecular (descriptor D_{CL}) relacionado con el reconocimiento topológico y la distancia entre grupos capaces de formar puentes de hidrógeno con el DNA, obtenido a partir de la longitud de la molécula geométricamente optimizada por métodos computacionales semiempíricos, teniendo en cuenta que se requiere una longitud específica para un reconocimiento geométrico óptimo con la doble hélice, considerando isohelicoicidad. [Chacón-García, 2001]

Un descriptor molecular es el resultado final de una lógica y de un procedimiento matemático que transforma la información química codificada dentro de una representación simbólica de una molécula en un número útil o el resultado de un cierto experimento estandarizado. [Todeschini, 2000] Los descriptores moleculares son números capaces de proveer datos teóricos y experimentales de la molécula. Aunque las propiedades moleculares no dependen solamente de la composición sino también de la conectividad, la molécula contiene implícita toda la información química, pero solamente una parte de ésta se puede extraer experimentalmente y el resto de manera teórica. El descriptor D_{CL} es un descriptor teórico que se obtiene a partir de la distancia de la molécula considerando los grupos extremos (NH_2-NH_2, o los que forman puentes de hidrógeno con el

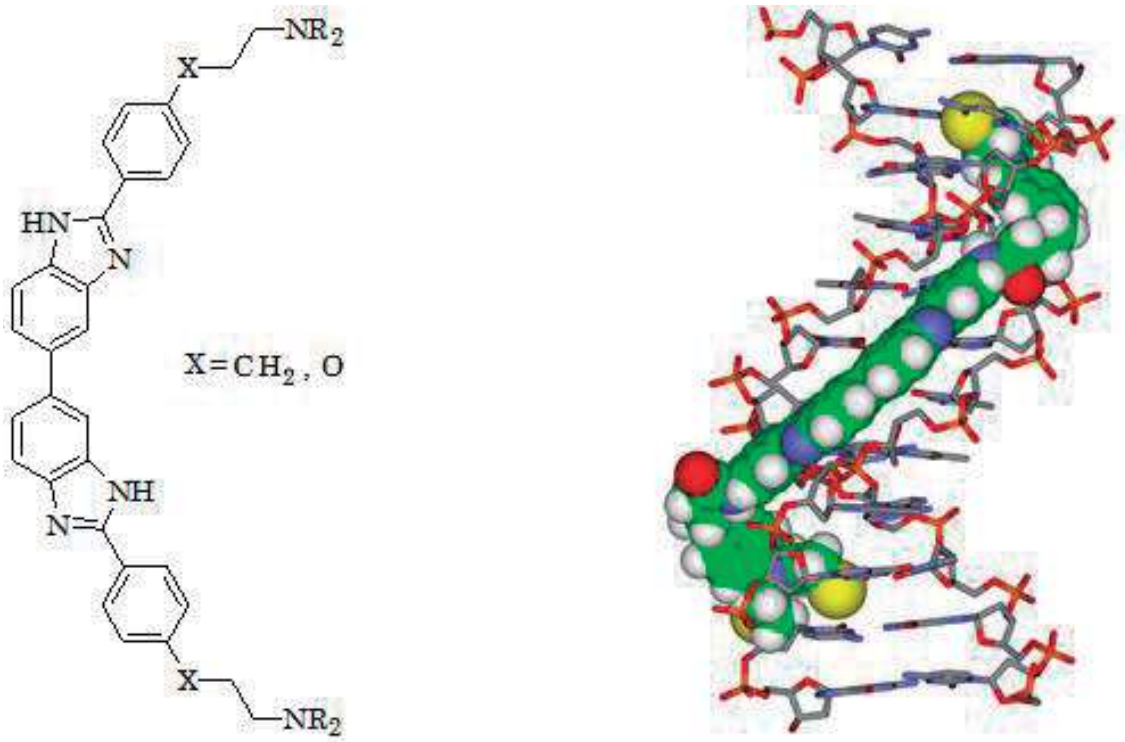


Figura 1. Representación de un reconocedor de surco y su interacción con el DNA

DNA), dividido por 4.202 o 3.4/cos 36. El resultado debe ser en principio el número de pares de bases más uno, que la molécula reconozca. Por lo tanto, si el número entero más cercano a este valor es restado al cociente de la división anterior, la diferencia que resulta será la separación del reconocimiento relativo, el cual es óptimo cuando es cero o inadecuado cuando es 0.5. [Chacón-García, 2001]

D_{CL}= |(Distancia entre NH_2-NH_2/4.202)– N| Ecuación 2.

Donde N es el número entero más cercano.

Dado lo anterior, resulta interesante explorar si el descriptor D_{CL} tiene aplicación a compuestos diferentes a los experimentados sobre todo con actividad biológica demostrada en DNA y así buscar compuestos novedosos con mayor actividad y menor toxicidad. Un parámetro de interacción de moléculas con el DNA es la Tm, es decir la temperatura a la cual un biopolímero se desnaturaliza en un 50%. Un incremento en la Tm representa mayor estabilidad del biopolímero. En el caso del DNA, la Tm se ve afectada al interactuar con intercaladores o RS, mismos que estabilizan la doble hélice.

En el presente trabajo se describe el estudio QSAR de una serie de reconocedores de surco con respecto a la Tm en DNA mediante Inteligencia Artificial utilizando AGs e incorporando en el método el descriptor D_{CL} diseñado ex profeso para el estudio de este tipo de compuestos.

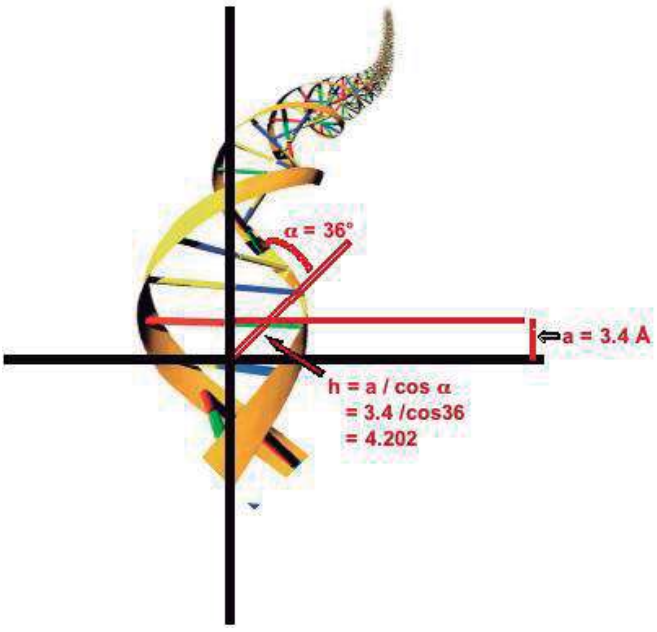


Figura 2. Esquema de las características del DNA que dan lugar al descriptor D_{CL}.

Materiales y métodos

La serie de 27 compuestos reconocedores de surco de DNA complementarios al dodecámero de Dickerson Drew (DDD) y la Tm que inducen en el DNA se obtuvieron de una base de datos descrita en la literatura. [De Oliveira, *et al.*, 2003] La optimización geométrica de cada estructura se llevó a cabo con el programa Hyper Chem. [Hyper Chem, v. 6.03 Professional, 2000] Las estructuras primeramente fueron modeladas por mecánica molecular y posteriormente se optimizó su geometría mediante el método semiempírico AM1 que es el más comúnmente utilizado en moléculas orgánicas pequeñas con las características estructurales de las aquí estudiadas. Para la optimización geométrica se tomó como base el algoritmo Polak-Riviere [Press, *et al.*, 1986], con parámetros de 0.1 kcal/(Å mol). Se procedió a obtener 1664 descriptores moleculares para cada uno de los compuestos, agrupados en 20 familias, con ayuda del programa DRAGON [Todeschini, *et al.*, 2005] y a estos se incorporó el descriptor D_{CL} que se obtuvo de acuerdo a su reporte en la literatura [Chacón-García, *et al.*, 2001]. Los valores de cada descriptor fueron importados al programa MobyDigs [Todeschini, *et al.*, 2004], por familias de descriptores y fueron "evolucionadas", utilizando Algoritmos Genéticos. La evolución se detuvo tomando como criterio 700 generaciones o bien cuando los primeros 50 modelos fuesen estables. Debido a que el programa solamente respalda 10 familias, el análisis se realizó en dos partes y se unieron cuando ambas llegan al criterio de parada. Cuando ambas partes alcanzaban las generaciones apropiadas, se evolucionaron juntas, para combinar los 50 mejores modelos de ambas y con ello obtener resultados más confiables.

Los mejores 50 modelos finales, fueron validados estadísticamente tomando como base R^2, Q^2, F y s, que arroja el programa mismo y que sirvieron de base para hacer predicción de Tm.

Utilizando Excel™ se hace la validación externa de cada modelo, utilizando el coeficiente de correlación y descartando aquellos modelos que mostraban correlación entre sus descriptores.

En el caso del coeficiente de correlación entre descriptores, se consideró como criterio un máximo de 0.5, excluyendo aquellos conjuntos que presentaron un valor superior a éste entre sus descriptores.

Con la finalidad de validar la capacidad predictiva del modelo, se extrajo de la serie de datos el compuesto H y se le calculó la actividad con los modelos obtenidos.

Resultados

Los mejores 50 modelos obtenidos después de 700 evoluciones en la primer etapa y aprox. 1200 en la segunda, se muestran en la **Tabla 1** en la que se aprecia que el descriptor D_{CL} fue arrojado en los conjuntos 39 y 48. Del conjunto 39 se obtiene el Modelo 1. (vide supra)

La matriz de correlación de los descriptores involucrados en los conjuntos 39 y 48 se presenta en la **Tabla 2**.

En el caso del conjunto 39, existe una correlación mayor a 0.5, entre los descriptores RDF095 y E1e, por lo que se fragmentó a tres descriptores explorando dos conjuntos nuevos descartando RDF095 y E1e en cada caso dando lugar a los Modelos 2 y 3.

Modelo 1:

log (ΔTm) = 0.0012 G(N..N) -0.3368 DCL + 0.0371
n = 27, r2 = 0.8349, F = 58.1512,
s = 0.0469, Q2 = 80.98

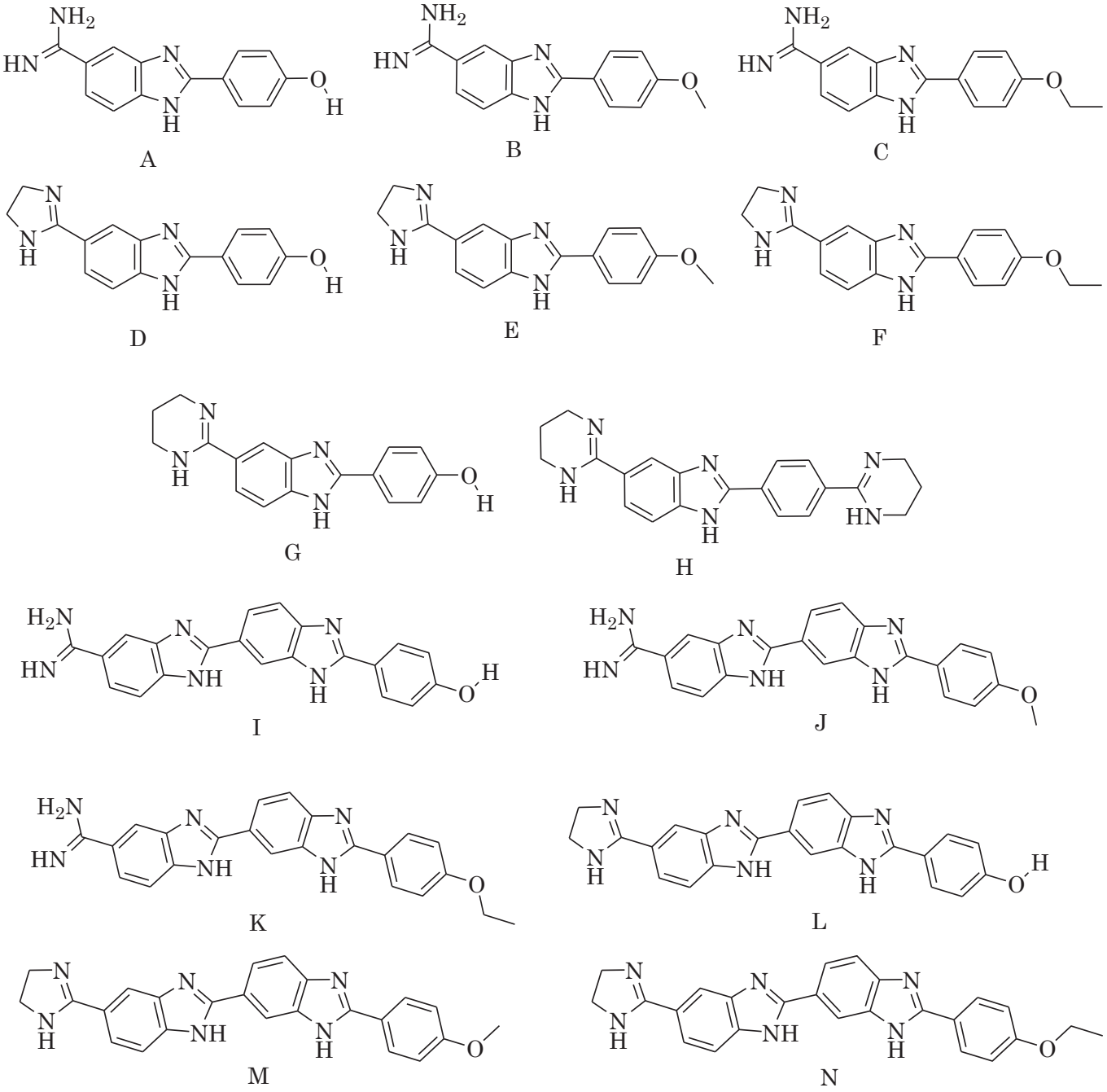


Figura 3. Estructra de los compuestos que conforman la serie de exploración.

Modelo 2:

log (ΔTm) = 0.0012 G(N..N) -0.0004 RDF095u -0.3374 DCL + 0.0403

n = 27, r2 = 0.8352, F = 37.1585, s = 0.0479, Q2= 80.79

Modelo 3:

log (ΔTm) = 0.0011 G(N..N) + 0.8935 E1e -0.3567 DCL -0.4731

n = 27, r2 = 0.8970, F = 63.8463, s = 0.0379, Q2= 87.57

La Tm calculada para cada compuesto se comparó con la experimental. Los valores de los descriptores, su respectiva Tm calculada y experimental se resume en la **Tabla 3**.

Discusión

La Tm está ligada de manera directamente proporcional a la estabilidad de la doble hélice del DNA y la interacción estabilizante de un fármaco con su blanco molecular incrementa sus posibilidades de coexistir en el ambiente celular. De ahí que entre mayor sea la Tm en los compuestos que interactúan con el material genético estos tendrán potencialmente actividad superior

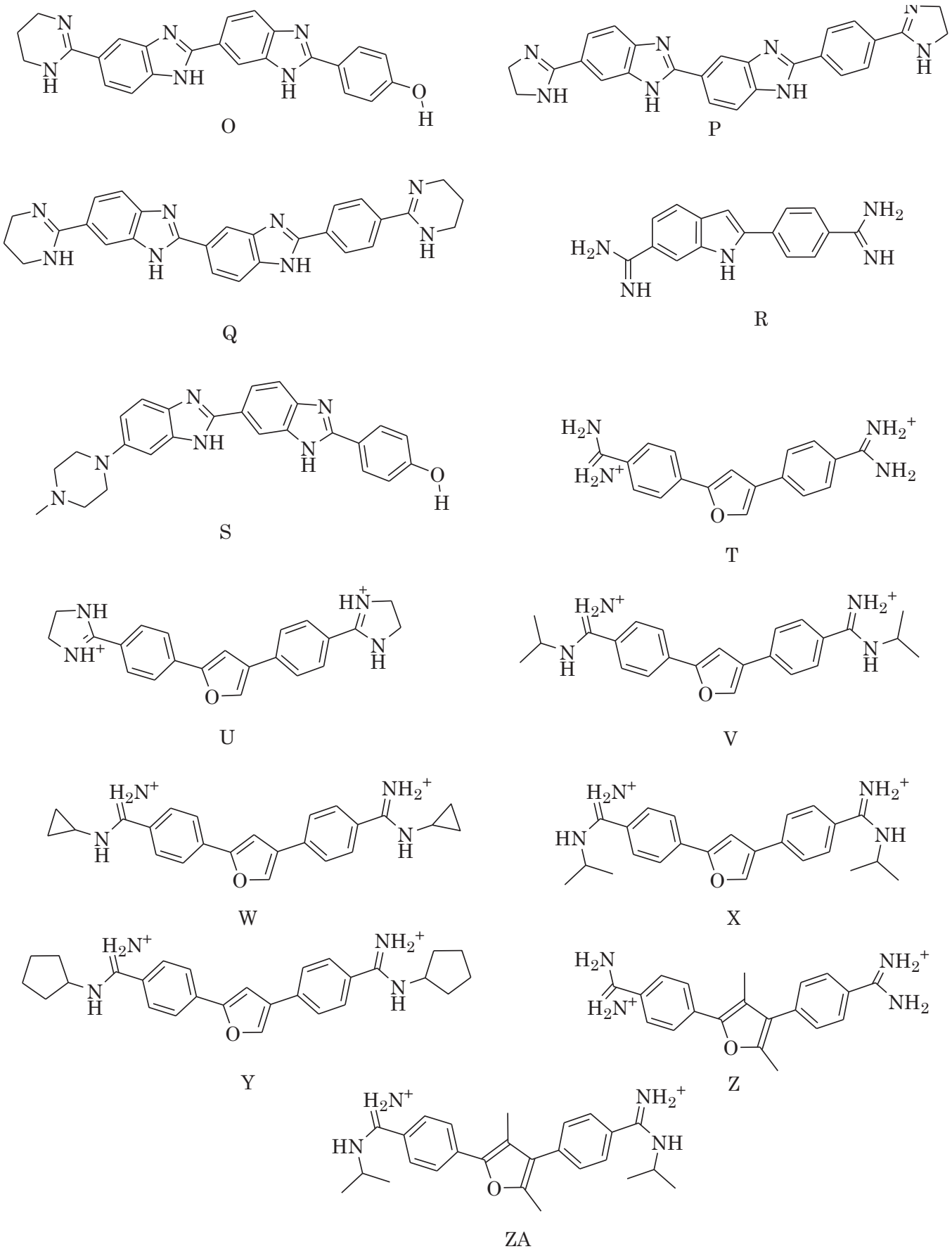


Figura 3. Estructra de los compuestos que conforman la serie de exploración (continuación).

Tabla 1. Modelos obtenidos con mejor ajuste estadístico

Conj. No.	Descriptores involucrados.	R^2	Conj. No.	Descriptores involucrados.	R^2
1	D/D ATS1m EEig02r GGI9	95.24	26	IDDE EEig02r GGI5 GGI8	94.94
2	Xu ATS1m EEig02r GGI9	95.51	27	EEig02r GGI5 GGI8	94.41
3	ATS2m MATS8e MATS8p GATS1p	95.35	28	EEig02r GGI5 GGI8 JGI3	94.8
4	EEig02r GGI4 GGI5 GGI8	95.32	29	piPC01 MATS8v MATS8e GATS1p	94.43
5	EEig02r GGI5 GGI8 LP1	95.27	30	EEig02r GGI5 GGI8 AEigZ	94.68
6	piPC01 MATS8e MATS8p GATS1p	95.13	31	EEig02r GGI5 GGI8 AEigm	94.68
7	ATS1m EEig02r GGI9 AEigm	95	32	EEig02r GGI5 GGI8 Eig1Z	94.68
8	ATS1m EEig02r GGI9 AEigZ	95	33	MATS8p EEig02r GGI5 GGI8	94.56
9	EEig02r GGI5 GGI8 H-048	95.14	34	EEig13x EEig02r GGI5 GGI8	94.51
10	EEig02r GGI5 GGI8 C-033	95.14	35	X0Av EEig02r GGI5 GGI8	94.68
11	ATS1m EEig02r GGI9 Eig1Z	94.99	36	VDA EEig02r GGI5 GGI8	94.67
12	EEig02r ESpm09u GGI5 GGI8	95.03	37	nC EEig02r GGI5 GGI8	94.56
13	EEig02r ESpm10u GGI5 GGI8	95.05	38	D/D EEig02r GGI5 GGI8	94.63
14	IVDM MATS8v MATS8e GATS1p	94.74	39	G(N..N) RDF095u E1e DCL1	93.26
15	EEig02r ESpm07u GGI5 GGI8	94.86	40	RBN EEig02r GGI5 GGI8	94.56
16	EEig02r GGI5 GGI8 E1e	95.12	41	GATS1p EEig02r GGI5 GGI8	94.55
17	EEig02r GGI5 GGI8 VEA1	94.98	42	X1v EEig02r GGI5 GGI8	94.53
18	EEig02r ESpm08u GGI5 GGI8	94.84	43	IDDM TIC0 GATS7v	92.78
19	EEig02r GGI3 GGI5 GGI8	94.97	44	IAC IDDM GATS7v	92.78
20	ATS1m MATS8v MATS8e GATS1p	94.56	45	GATS7v G(N..N) Ui	87.02
21	ATS1m MATS8e MATS8p GATS1p	94.72	46	GATS7v EEig02r	86.42
22	EEig02r GGI5 GGI8 C-026	94.8	47	E1u nCbH C-002 Ui	89.15
23	EEig02r GGI5 GGI8 C-040	94.8	48	G(N..N) DCL1	83.49
24	EEig02r GGI5 GGI8 E1u	95.04	49	GATS7v Ui	84.58
25	EEig02r GGI5 GGI8 nCconj	94.81	50	nCar C-033 Ui	84.11

lo cual, claro está, sin considerar otros factores como absorción o metabolismo del fármaco. La importancia del descriptor D_{CL} radica justamente en este punto.

A partir de un análisis de regresión lineal simple, la relación entre D_{CL} y la Tm no resulta estadísticamente significativa por lo que se infiere que la geometría del reconocimiento por puentes de hidrógeno en función de su helicoicidad, como único factor a considerar, no es suficiente para explicar este parámetro termodinámico. Sin embargo, en los Modelos 1 a 3 se aprecia que tanto R^2, F, s, y Q^2 son estadísticamente significativos y que al menos considerando los 1665 descriptores involucrados en la búsqueda de modelos, que permitan explicar y predecir la Tm de los reconocedores de surco considerados en este estudio, hay otros factores que pueden afectar el reconocimiento y la estabilidad de los compuestos de la serie explorada y el DNA.

Tabla 2. Correlación entre los descriptores involucrados en los modelos 1- 3.

	G(N..N)	RDF095u	E1e	DCL
G(N..N)	1	0.44	0.09	-0.23
RDF095u		1	0.53	-0.15
E1e			1	0.1
DCL				1

En el caso del Modelo 1, con base al coeficiente de determinación se establece que el 83.5% de la variabilidad de Tm es explicada tanto por el descriptor D_{CL} como por el descriptor geométrico G(N..N) que considera la distancia entre nitrógenos, de existir dos, lo cual resulta razonable pues el adecuado reconocimiento geométrico es esencial en la formación de la supramolécula. A diferencia del Descriptor D_{CL}, el descriptor G(N..N) no involucra la helicoicidad del DNA y la distancia entre nitrógenos se obtiene a partir de la estructura geométricamente optimizada al vacio que no necesariamente corresponde a la conformación en su sitio de acción.

El Modelo 2 involucra, además de los descriptores que participan en el Modelo 1, el descriptor teórico RDF095u; que forma parte de los descriptores de función de distribución radial (del cual derivan sus siglas: Radial Distribution Function).

El término exponencial incluido en la ecuación que define la función RDF es (R - rij)^2, donde rij es la distancia entre el

Tabla 3. Comparación de Tm calculada y experimental y valores de los descriptores moleculares utilizados en los modelos 1-3.

Valores de Descriptores					Modelo 1		Modelo 2		Modelo 3	
Mol	G(N..N)	E1e	RDF095u	DCL	Tm experimental	Tm calculada	Tm experimental	Tm calculada	Tm experimental	Tm calculada
A	24.316	0.575	4.758	0.488	0.845	0.81	0.845	0.7994	0.845	0.783
B	23.889	0.562	7.728	0.406	0.792	0.813	0.792	0.8488	0.792	0.814
C	24.317	0.554	7.057	0.499	0.69	0.812	0.69	0.791	0.69	0.743
D	24.437	0.574	4.632	0.476	0.748	1.399	0.748	0.8072	0.748	0.789
E	24.514	0.557	6.465	0.467	0.875	1.263	0.875	0.8118	0.875	0.768
F	24.571	0.556	2.24	0.462	0.845	1.399	0.845	0.8181	0.845	0.77
G	24.118	0.58	6.573	0.462	0.845	1.382	0.845	0.814	0.845	0.808
I	95.084	0.578	11.616	0.009	1.362	1.391	1.362	1.3996	1.362	1.4
J	95.192	0.582	8.696	0.141	1.362	1.339	1.362	1.2668	1.362	1.267
K	95.169	0.567	13.052	0.0088	1.362	1.25	1.362	1.3984	1.362	1.369
L	95.612	0.582	11.051	0.026	1.378	1.54	1.378	1.3839	1.378	1.394
M	95.454	0.584	16.137	0.018	1.387	1.487	1.387	1.3857	1.387	1.408
N	95.604	0.573	11.548	0.067	1.412	1.118	1.412	1.3399	1.412	1.323
O	94.905	0.587	12.857	0.153	1.405	1.346	1.405	1.2495	1.405	1.266
P	250.12	0.579	15.269	0.426	1.479	0.976	1.479	1.5434	1.479	1.482
Q	248.159	0.595	17.576	0.464	1.487	0.994	1.487	1.4874	1.487	1.477
R	78.287	0.566	5.165	0.239	1.233	0.969	1.233	1.1243	1.233	1.083
S	98.504	0.583	11.24	0.071	1.253	0.884	1.253	1.3468	1.253	1.356
T	59.421	0.584	7.935	0.349	0.959	0.868	0.959	0.9778	0.959	0.978
U	59.678	0.584	8.019	0.326	0.857	0.951	0.857	0.996	0.857	0.998
V	59.322	0.621	12.935	0.358	1.037	0.974	1.037	0.9664	1.037	1.048
W	59.06	0.657	15.291	0.475	1.037	0.946	1.037	0.88	1.037	1.024
X	58.745	0.611	12.924	0.498	0.968	1.352	0.968	0.8655	0.968	0.913
Y	59.085	0.679	22.585	0.381	1.113	0.81	1.113	0.9406	1.113	1.157
Z	59.719	0.524	8.178	0.353	0.813	0.813	0.813	0.9753	0.813	0.862
ZA	59.311	0.565	25.011	0.388	0.732	0.812	0.732	0.9341	0.732	0.911
H	105.089	0.599	12.542	0.088	1.233	1.399	1.233	1.3518	1.233	1.405

par de átomos i.j y R es la distancia de referencia en intervalos de 1.0 a 15.5 Ang. Esto significa que todos los pares de átomos con distancias geométricas cerca de R juegan un papel significante en el descriptor. Este descriptor puede estar ponderado por masas atómicas, electronegatividad, polarizabilidad, etc. En este caso, RDF095u, establece a R como 9.5 Ang y la sigla u indica que no está ponderado por ningún término en particular (unweighted). Esta información se resume a que las distancias geométricas dentro de la estructura en exactamente 9.5 Ang juegan un papel importante en el reconocimiento molecular y por lo tanto en la estabilidad del complejo DNA-Reconocedor de surco formado. Del Modelo 2, se aprecia que RDF095u tiene una contribución negativa y por lo tanto entre mayor es este componente numéricamente, contribuye disminuyendo a la Tm. Sin embargo, el coeficiente de RDF095u en el modelo es de 0.0004, lo cual se traduce a una contribución matemáticamente insignificativa y los datos calculados a partir del modelo con respecto a los experimentales son prácticamente iguales a los que respecta al Modelo 1.

Por su parte, el Modelo 3, que al igual que el Modelo 2 tiene el mismo origen en su conjunto de datos obtenido por el proceso de evolución, incluye al descriptor E1e que forma parte de los descriptores conocidos como WHIM (weighted holistic invariant molecular descriptors). Estos son índices moleculares 3D que representan diferentes fuentes de información química en términos de tamaño, forma, simetría y distribución atómica y son obtenidos a partir de sus coordenadas de la estructura tridimensional con diferentes esquemas de ponderación. El descriptor E1e está ponderado en las electronegatividades de Sanderson (que se fundamenta en el reciproco del volumen atómico). Es interesante este resultado pues la Tm está definida, como se mencionó anteriormente, por interacciones estabilizantes dentro de las cuales se encuentran las interacciones dipolo-dipolo y estas a su vez están relacionadas directamente con la diferencia

de electronegatividad de las especies que interactúan. En este caso la variabilidad del modelo del 89.7 % es explicada por los descriptores D_{CL}, G(N..N) y E1e, es decir por aspectos de helicoicidad, geométricos y de electronegatividad.

Aunque la capacidad predictiva de los modelos 1 a 3 es alta (Q^2 = 80.98, 80.79, 91.31, respectivamente) y la validación externa con el compuesto **H** así lo demuestra, el análisis de regresión mostró ser ligeramente superior en el Modelo 3, mismo que explica de manera más amplia la interacción con el DNA no solo considerando aspectos geométricos sino además electrónicos. La contribución del descriptor D_{CL} es importante en los tres modelos y abre la posibilidad de utilizarse, en conjunto con el descriptor E1e y G(N..N) para predecir la Tm de compuestos potencialmente reconocedores de surco que no han sido explorados en QSAR, por ejemplo compuestos peptídicos.

Agradecimientos

Con apoyo económico de CIC proyecto 2.18

Referencias

Bewley, Carole A., .Gronenborn Angela M. y Clore Marius G. *(1998). Minor Groove – Binding. Architectural proteins: Structure,* Funtion, and DNA Recognition. *Annual Review of Biophysic and Biomolecular Struct*ure 27: 105-31.

Chacón-García Luis, Martínez Roberto.(2001) Cytotoxic activity and QSAR of N,N' –diarylalkanediamides. *European Journal of Medicinal Chemistry* 36: 731-736.

De Oliveira, A. M.; Custodio, F. *B.; Donnici,* C. L.; Montanari, C. A. (2003) *European Journal of Medical Chemistry* 38: 141-155.

Estévez Valencia Pablo. (1997). Optimización mediante algoritmos genéticos. *Anales del Instituto de Ingenieros de Chile. Agosto. pp. 83-92.http://www.inele.ufro.cl/apuntes/Tutores_Inteligentes/Pablo%20Estevez%20-%20Optimizacion%20Mediante%20AGs.pdf*

Hou, T.J., Wang J.M. *y Xu X.J. (1999). Aplication of genetic algorithms on the structure – activity correlation study of a group of non-nucleoside HIV – 1 inhibitors. Chemometrics and intelligent laboratory systems 45: 303-310.*

Hyper Chem 6.03 for Windows; HyperCube Inc., 2000.

Koga Hiroshi, Itoh Akira, Murayama Satoshi, Suzue Seigo y Irikura Tsutomu. (1980). Structure- Activity Relationships of Antibacterial 6,7-and 7,8- Disubstituted 1-Alkyl-1,4-dihydro-4-oxoquinoline-3-carboxylic Acids. *European Journal of Medical Chemistry 23: 1358-1363.*

Larrañaga Pedro e Inza Iñaki. Tema 2. Algoritmos genéticos. *Departamento de Ciencias de la computación e Inteligencia Artificial. Universidad del País Vasco – Euskal Herriko Unibertsitatea. http://www.redksr.com/~talos/t2geneticos.pdf*

Martínez R. y Chacón-García L. (2005) *Current Opinion in Medicinal Chemistry. 12: 127-151.*

Neidle Stephen. (2001). DNA minor-groove recognition by small molecules. *Natural Product Reports 18: 291-309.*

Pastor M. y Alvarez – Builla J. Técnicas QSAR en diseño de fármacos. *Departamento de Química Orgánica. Universidad de Alcalá. E-28871 Alcalá de Henares. Madrid.*

Press W, H., et. al., *(1968). Numerical Recipes: The Art of Scientific Computing,* Cambridge University Press, Chapter 10.

Software Microsoft Office Excel 2007. Para Windows. Microsoft Corporation. *2006.*

Stefanic-Petek Alenka, Krbavcic Ales, Solmajer Tom. (2002). QSAR of flavonoids: differential inhibition of aldose reductase and p56lck protein tyrosine kinase. *Croatica Chemica Acta. CCACAA 75 (2) 517-529.*

Suh Myung-Eun, Park So-Young y Lee Hyun-Jung. (2002). Comparison of QSAR Methods of Anticancer 1-N-Substituted Imidazoquinoline-4,9-dione Derivatives. *Bull. Korean Chemical. Society. Vol. 23,* No. 3, 417

Todeschini Roberto y Consonni V., (2000). Manual de descriptores moleculares, Wiley-VCH. *http://www.moleculardescriptors.eu/books/handbook.htm*

Todeschini R.; Ballabio, D.; Consonni, V.; *Mauri,* A.; Pavan, M. (2004) Mobydigs Computer Software, 1.0; TALETE srl: Milano.

Todeschini R., Consonni, V.; Mauri, A.; *Pavan,* M. (2005). DRAGON—Software for the calculation of molecular descriptors. Ver. 5.3, Talete srl, Milano, Italy.

Turabekova Malakhat A. y Rasulev Bakhtiyor F. (2004). *A QSAR Toxicity of a Series of Alkaloids with the Lycoctonine Skeleton. Molecules 9: 1194-1207.*

Yalcin Ismail, Ören Ilkay, Temiz Özlem y Aki Sener Esin. (2000). QSARs of some novel isosteric heterocyclics with antifungal activity. *Acta Biochimica Polonica. 2: 481-486.*