Modelos bioinformáticos y estudio de receptores de proteínas mediante el uso de redes complejas para el desarrollo y diseño de fármacos eficaces en patologías del sistema nervioso central.

Memoria presentada por

Manuel Quintín Escobar Cubiella

Para optar al grado de
Doctor en Farmacia

Departamento de Química Orgánica
Facultade de Farmacia

Directores:

Dr. Francisco Prado Prado
Prof. Dr. Xerardo García Mera
Dr. Humberto González Díaz

Santiago de Compostela, Abril 2012

D. Xerardo García Mera, Prof. Titular y D. Francisco Javier Prado Prado, PDI Doctor Contratado por el Programa Ángeles Albariño, ambos del Departamento de Química Orgánica de la Universidad de Santiago de Compostela (USC), así como D. Humberto González Díaz, Doctor Contratado por el Departamento de Microbiología y Parasitología, Área de Parasitología, Facultad de Farmacia, USC.

CERTIFICAN:

Que la memoria titulada: "*MODELOS BIOINFORMÁTICOS Y ESTUDIO DE RECEPTORES DE PROTEÍNAS MEDIANTE EL USO DE REDES COMPLEJAS PARA EL DESARROLLO Y DISEÑO DE FÁRMACOS EFICACES EN PATOLOGÍAS DEL SISTEMA NERVIOSO CENTRAL*", que para optar al grado de Doctor Farmacia presenta MANUEL QUINTÍN ESCOBAR CUBIELLA, ha sido realizada bajo nuestra dirección, en el Departamento de Química Orgánica de la Facultad de Farmacia de la Universidad de Santiago de Compostela.

Y considerando que el trabajo constituye tema de Tesis Doctoral, autorizamos su presentación en la Universidad de Santiago de Compostela.

Y para que conste, expedimos el presente certificado en Santiago de Compostela a nueve de abril del dos mil doce.

_____ _____

Fdo.: Prof. Dr. Xerardo García Mera Fdo.: Dr. Francisco Javier Prado Prado

_____

Fdo.: Dr. Humberto González Díaz

Castro soy y Castro he sido

Asiento en firme Montaña
Y a la Corona de España
Con lealtad siempre he servido
Armas, Escudo y Señal
Castillo, Puente y Santa Ana
Naos, Ballena y mar llana

Son de Castro la Leal

Divisa de la Ciudad de Castro-Urdiales.

Siglo XIII-XIV.

Agradecimientos

Hace 20 años ya, que comenzó mi relación con la Universidad de Santiago de Compostela, y puedo afirmar, que desde el primer momento se afianzó en mi la ambición por el saber y el conocimiento. En primer lugar, tuve la enorme suerte de vivir mis experiencias universitarias en un entorno multicultural, el C.M.U. Gelmírez, donde me embadurné de amigos y compañeros geniales, de orígenes e inquietudes tan dispares con los que las tertulias y disputas dialécticas más alocadas se hacían sabrosamente interminables. En la Facultad de Farmacia fui descubriendo diversos aspectos de la profesión que aún sigo aplicando a día de hoy, pero especialmente me quedé enganchado de la asignatura de Química Farmacéutica, a mi entender, una de las disciplinas más completas de la carrera, y cuando ya había leído la Tesina y realizado los cursos de Doctorado, entro en contacto con la Industria Farmacéutica, teniendo el privilegio de haber aprendido y trabajado en grandes compañías, Janssen-Cilag, Merck Sharp & Dohme y los últimos 8 años en Novartis Pharmaceutica, y mira por donde mi camino se vuelve a cruzar con la USC a fin de rematar otro paso con la inquietud y ambición de no permitir que sea el último.

Quiero dedicar esta memoria a las personas que con su esfuerzo y sacrificio han facilitado que me pudiera mover por el mundo adelante sin más preocupación que la de buscar mi sitio. Gracias a mis padres Manolo y Rebeca por su ayuda constante e incondicional y por el ejemplo que siempre he recibido de vosotros.

A Cristina quiero agradecerle su comprensión por el tiempo robado y por el amor que me ha dedicado siempre, confiando ciegamente en mis posibilidades, además de por los dos hijos más guapos del mundo, Mikel y Helena, que algún día seguirán sus propias inquietudes. Espero saber estar a la altura….

A mi hermano Javier, primer Doctor de la Familia, porque me ha servido de acicate.

A Xerardo García Mera, que desde el año 1994 me ha espoleado y prestado su modelo en la forma de entender la Ciencia.

A Fran, por hacerme llegar a entender conceptos que ni sabía que existían, gracias por tu paciencia para conmigo.

A Humber por su inestimable colaboración.

Abreviaturas utilizadas

LDA: Linear **D**iscriminant **A**nalysis, término que proviene del inglés: Análisis Discriminante Lineal.

AChE: Acetilcolinesterasa.

ANN: Artificial **N**eural **N**etworks, término que proviene del inglés: Redes neuronales artificiales.

3D: Tridimensional.

CM: Cadenas de **M**arkov.

DTP: Pares de fármaco-proteína con alta afinidad.

nDTP: Pares de fármaco-proteína con nula afinidad.

EA: Enfermedad de Alzheimer.

EP: Enfermedad de Parkinson.

FDA: Food and **D**rug **A**dministration of USA; Administración de alimentos y medicamentos de los EE.UU.

HTS: **H**igh-**T**hroughput-**S**creening, término que proviene del inglés: evaluación de alta eficacia.

IT: Índices topológicos

LNN: Lineal **N**eural **N**etwork, término que proviene del inglés: Red neuronal lineal.

MARCH-INSIDE (MI): Markov **Ch**ain **I**nvariants for **N**etwork **Si**mulation and **De**sign

PDB: Protein **D**ata **B**ank; Banco de datos de proteínas.

QSAR: Quantitative-**S**tructure-**A**ctivity-**R**elationship, término que proviene del inglés: relación-cuantitativa-estructura-actividad.

mt-QSAR: multi-target **Q**uantitative-**S**tructure-**A**ctivity-**R**elationship, término que proviene del inglés: relación-cuantitativa-estructura-actividad multi-target.

QSPR: Quantitative-**S**tructure-**P**roperty-**R**elationship, término que proviene del inglés: relación-cuantitativa-estructura-propiedad.

QSTR: Quantitative-**S**tructure-**T**oxicity-**R**elationship, término que proviene del inglés: relación-cuantitativa-estructura-toxicidad.

SN: Sistema **N**ervioso.

Índice

1. Introducción

El Sistema Nervioso (SN) es el sistema orgánico más complejo e importante del cuerpo humano, consta de varias estructuras altamente especializadas que coordinan y activan múltiples tareas del organismo. Es una unidad tan compacta que la lesión o degeneración de cualquiera de sus estructuras provoca daños considerables.

Una de las patologías más relevantes que afecta al SN es la Enfermedad de Parkinson (EP), que cursa de forma progresiva y debilitante generando un grave deterioro de la capacidad motora del paciente provocando un tremendo impacto en el paciente y su entorno. Podemos afirmar que en la actualidad no existe ningún tratamiento capaz de curar la EP y su pronóstico a largo plazo no es nada halagüeño. Los tratamientos actuales están enfocados en la mejora de la calidad de vida del paciente paliando los síntomas de la enfermedad, pero incluso los mejores fármacos no están exentos de reacciones adversas y de una pérdida de eficacia con el paso del tiempo. Hoy más que nunca resulta crucial encontrar nuevas terapias ante una patología con un factor de riesgo incuestionable como es la edad del individuo[1, 2].

La mayor parte de los casos nuevos de EP se dan en pacientes entre 50-70 años y la prevalencia actual en España es de 1 por cada 100 habitantes y se espera un incremento considerable de este dato debido al aumento de la esperanza de vida de la población.

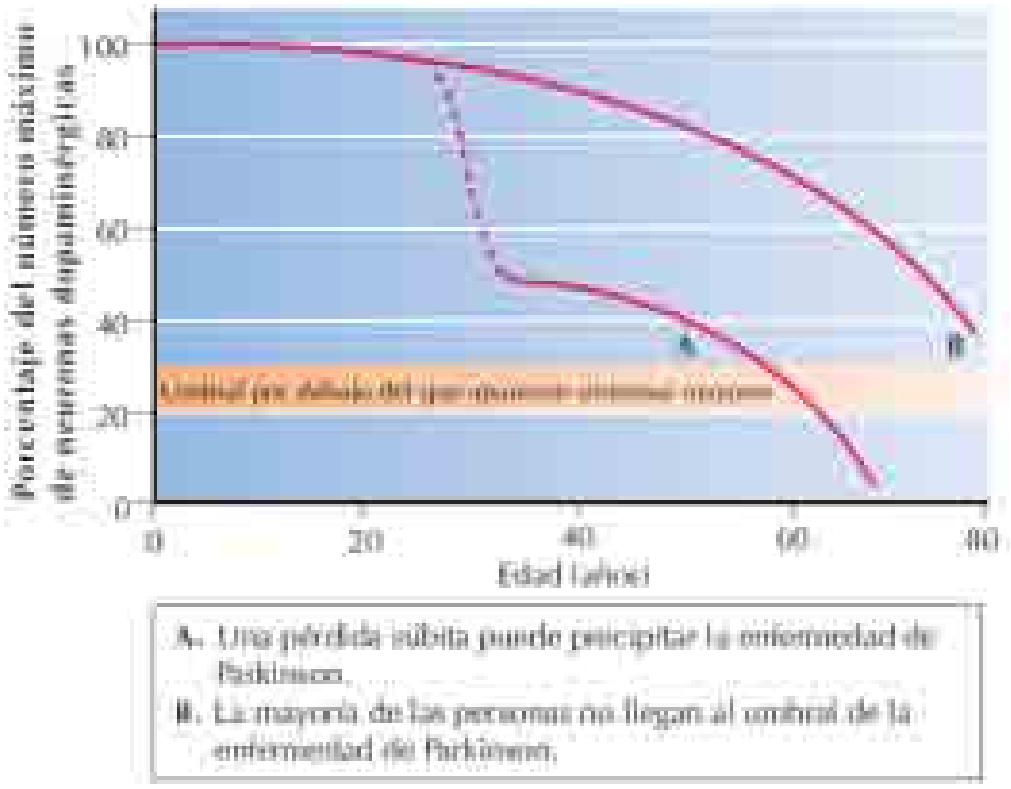
La incidencia en nuestro país es de 16 casos nuevos de EP por cada 100.000 habitantes y es de destacar que la mortalidad asociada ha disminuido considerablemente en las últimas tres décadas en pacientes mayores de 65 años, fundamentalmente debido a la introducción del tratamiento con levodopa, que ha aumentado en 5 años la media de edad a la que mueren los pacientes con EP.

El diagnóstico temprano es complicado ya que los síntomas aparecen de una forma lenta y gradual, pero una vez establecida la enfermedad los signos anatomopatológicos son concluyentes: Pérdida neuronal generalizada en la sustancia negra, disminución de los niveles de dopamina (DA) y presencia de cuerpos de Lewy.[a]

[a] *Cuerpos de Lewy*: depósitos anómalos que se observan en el interior de las neuronas y que parecen estar asociados a la enfermedad de Parkinson y a otras enfermedades neurológicas; el número de cuerpos de Lewy parece aumentar a medida que empeora la enfermedad.

Cuando, por causa desconocida, disminuye el número de neuronas dopaminérgicas de la sustancia negra, disminuye la concentración de DA, produciéndose un deterioro en el equilibrio de los distintos neurotransmisores y cuando esos niveles de DA disminuyen en un 70-80% aparecen los síntomas motores tan característicos de la EP: bradicinesia (lentitud en los movimientos voluntarios), acinesia (ausencia de movimiento), temblores y rigidez muscular [3,4].

Relación entre la pérdida de neuronas dopaminérgicas,
la edad y la aparición de los síntomas



A. Una pérdida súbita puede precipitar la enfermedad de Parkinson.
B. La mayoría de las personas no llegan al umbral de la enfermedad de Parkinson.

La palabra demencia describe un grupo de síntomas causados por cambios en la función cerebral, viéndose disminuidas ciertas capacidades como el almacenamiento de la información y con una marcada pérdida de la memoria.

Las demencias son síndromes adquiridos que se manifiestan con signos y síntomas de deterioro persistente de las capacidades cognitivas. Alteran la capacidad funcional en el ámbito social, laboral y familiar provocando un deterioro y cambios en la personalidad del individuo.

Las causas que pueden provocarlas son variadas, algunas son reversibles y otras no. Los dos tipos de demencia más frecuentes son la demencia secundaria asociada a la

Enfermedad de Alzheimer (EA) y la demencia vascular o demencia multi-infarto, provocada por una falta de irrigación sanguínea en el cerebro.

La disminución normal de las capacidades mentales asociadas al envejecimiento consiste sobre todo en cambios leves en la memoria y en la velocidad de procesamiento de la información; estos cambios no son progresivos y habitualmente no afectan al funcionamiento diario de la persona. El envejecimiento normal se asocia con una disminución en la adquisición de nuevos conocimientos y también en la memoria de hechos recientes, sin embargo los recuerdos antiguos no se alteran[5].

Hay que tener en cuenta que la EA no forma parte del envejecimiento normal, su prevalencia es inferior al 1% en pacientes entre 60-64 años y se incrementa de forma exponencial con la edad, llegando a una prevalencia entre 24-33% en mayores de 85 años.

La EA es un trastorno neurológico progresivo, de origen desconocido, que provoca la muerte de las células nerviosas en el cerebro. A medida que la enfermedad progresa se deteriora la cognición y pueden surgir modificaciones de la personalidad y alteraciones conductuales. En su etapa avanzada, la EA conduce a la demencia y finalmente a la muerte.

La EA es, junto con otras enfermedades demenciales, la 4ª causa principal de muerte en España en personas mayores de 65 años. Tiene un gran impacto económico y social, convirtiendo esta patología en un problema de primer orden para familiares y cuidadores debido al estrés que genera el deterioro progresivo en las funciones físicas y mentales del enfermo[6].

El origen de la EA es desconocido en un 90% de los pacientes, los cuales desarrollan la enfermedad entre los 35-55 años. El 10% restante es de origen hereditario y suele desarrollarse a partir de los 60 años.

Se ha podido comprobar que la disminución de la función colinérgica se correlaciona con el déficit cognitivo de los pacientes y que las neuronas colinérgicas, críticas para la memoria y el aprendizaje, muestran cambios degenerativos en la EA. Sin embargo, el proceso primario por el que se desarrolla la EA aún permanece desconocido. Por eso se piensa que los fármacos que mejoran esta función colinérgica, tanto agonistas colinérgicos como inhibidores de las enzimas que metabolizan este neurotransmisor, pueden interferir positivamente en la progresión de la EA y mejorar el estado cognitivo de los pacientes [7].

La búsqueda y desarrollo de fármacos eficaces para el tratamiento de estas enfermedades ha generado grandes expectativas, debido a la relevancia que tienen sobre la economía de los sistemas sanitarios y la tremenda carga y desgaste que sufren familia y cuidadores. Por ello, la industria farmacéutica se ha volcado sobre estas patologías en las últimas tres décadas, pero las dificultades de realizar ensayos sobre el SN provoca que los gastos y tiempos de investigación se disparen, limitando de forma considerable la rentabilidad de los procesos tradicionales en el desarrollo de nuevos medicamentos[8].

Es en este apartado donde realiza sus aportaciones el diseño de fármacos, dedicando una parte del mismo al desarrollo de modelos matemáticos que permitan predecir propiedades de interés para una gran variedad de sistemas químicos incluyendo moléculas de bajo peso molecular, polímeros, biopolímeros, sistemas heterogéneos, formulaciones farmacéuticas, conglomerados de moléculas e iones, materiales, nano-estructuras y otros.[9] Este tipo de predicciones tienen como objetivo fundamental complementar y evolucionar las técnicas de carácter experimental tradicionales, fundamentalmente colaborando en la obtención de nuevas moléculas activas con mayor probabilidad de éxito, con la ventaja que de ello se deriva en términos de ahorro de tiempo, recursos materiales y también en el refinamiento y reducción en el uso de animales de laboratorio.[10-12]

En la actualidad existen miles de compuestos químicos, ya sea de origen natural o de síntesis, y en su amplia mayoría aún no se les ha encontrado aplicaciones farmacológicas, agroquímicas, industriales o de algún otro tipo. Esto es consecuencia directa de la gran diferencia que existe entre la velocidad con que los nuevos compuestos son obtenidos y caracterizados en la mesa de laboratorio y la realización de los ensayos experimentales que permitan evaluar su potencial terapéutico.

Cuando estos productos pretenden ser destinados al consumo humano en aras de conseguir mejoras terapéuticas, los procesos aún sufren una ralentización adicional por la cantidad de trámites administrativos y burocráticos destinados a garantizar que los fármacos lleguen al consumo humano con todas las garantías necesarias para los pacientes. Por otra parte existen determinadas patologías en las cuales los ensayos de laboratorio son de una gran complejidad en sí mismos, como ocurre en el caso de los compuestos con potencial actividad antiviral, o bien los ensayos son muy costosos tanto en términos de recursos materiales, humanos y de tiempo, como ocurre con los compuestos con potencial actividad neuroprotectora dirigidos al tratamiento de diferentes enfermedades degenerativas como EA y EP.

Este motivo ha obligado a la industria farmacéutica a cambiar sus estrategias de búsqueda enfocando sus esfuerzos hacia el desarrollo de métodos que racionalicen los sistemas de diseño y evaluación de nuevos fármacos ya en las primeras fases del descubrimiento de los mismos, produciéndose así una importante disminución del coste económico y temporal en el desarrollo de nuevos compuestos para su uso farmacéutico. Por ello, y en los últimos años, ha ido ganando una gran importancia el desarrollo de modelos capaces de predecir las rutas de descubrimiento con mayor probabilidad de éxito y que sirvan de guía al investigador en el desarrollo racional de fármacos con un importante ahorro de recursos.

En dicho sentido, los estudios QSAR (Quantitative Structure-Activity-Relationships) son usados cada vez mas como herramientas para el descubrimiento molecular. Estos modelos QSAR pueden ser diseñados para que predigan la probabilidad de que un fármaco sea efectivo contra una enfermedad degenerativa determinada ya sea la enfermedad de Parkinson, Alzheimer o cualquier otra, actuando sobre una diana molecular específica. Para ello, primeramente deberán recopilarse de las bases de datos públicas los datos de actividad biológica de fármacos neuroprotectores con diana molecular conocida. Seguidamente, se calcularán determinados parámetros numéricos llamados Índices Topológicos (ITs) tanto de los fármacos como de sus dianas moleculares[13] y posteriormente por análisis estadístico y/o Inteligencia Artificial (Redes Neuronales Artificiales) se buscarán los modelos QSAR. Dichos ITs describen únicamente la topología ó conectividad en fármacos y secuencias de sus dianas (proteínas y/o ADN/ARN) y, por ello, son versátiles y fáciles de usar. Además se pueden explorar grandes bases de datos con un importante ahorro de tiempo y recursos materiales[14].

Estos modelos duales QSAR Bioinformáticos + Quimioinformáticos son de interés para los grupos de investigación dedicados a la Química Farmacéutica. Se pueden usar estos modelos para predecir la probabilidad de actuar como dianas de fármacos a nuevas proteínas y/o ADN/ARN que se aíslen y que actúen dentro del desarrollo de las enfermedades demenciales, así como otras ya conocidas pero con función desconocida como diana de fármacos.

En esta memoria presentamos de manera conjunta la revisión de modelos previos y trabajos específicos novedosos, en los que se han introducido nuevos índices numéricos utilizados para describir tanto la estructura molecular de fármacos como la estructura macromolecular de sus dianas o receptores (proteínas y/o ADN/ARN). Con

estos ITs hemos sido capaces de desarrollar nuevos modelos multiQSAR de gran interés por su doble función en la predicción de fármacos y sus dianas moleculares. Estos trabajos permitirán la introducción de nuevos conceptos teóricos y la evolución hacia modelos con posibles aplicaciones en la búsqueda de nuevos fármacos neuroprotectores útiles en el tratamiento de las enfermedades de Parkinson y Alzheimer y/o nuevas dianas moleculares para estos fármacos. Este tipo de investigación abarca un área general-básica en la que interactúan la Bioinformática y la Quimioinformática.

1.1 Metodología QSAR

Esta metodología se basa en el uso de cálculos por ordenador y en las nuevas tecnologías de la informática las cuales pueden ser usadas tanto para pequeñas moléculas como para macromoléculas.

Para moléculas pequeñas:

1. Estudios de relación cuantitativa estructura molecular-actividad farmacológica (QSAR) y de estructura molecular-propiedades toxicológicas y eco-toxicológicas incluyendo mutagenicidad y carcinogénesis (QSTR).
2. Predicción de propiedades químicas y fisicoquímicas de moléculas. Estudios de relación estructura molecular y propiedades de absorción, distribución, metabolismo y eliminación (ADME).
3. Predicción de mecanismos de acción biológica de moléculas y evaluación *in silico* de alta eficacia para grandes bases de datos (virtual HTS).

Para macromoléculas:

4. Estudios de interacción fármaco-receptor (neuronas).
5. Bioinformática aplicada a estudios de relación secuencia-función y propiedades estructurales de ácidos nucleicos y proteínas.
6. Búsqueda de nuevas dianas terapéuticas y "sitio activo" a partir de datos de Genómica y/o Proteómica.
7. Búsqueda de biomarcadores para diagnóstico de enfermedades o como indicadores de contaminación.
8. Predicción de propiedades fisicoquímicas de polímeros sintéticos, biopolímeros, materiales y nano-estructuras.
9. Predicción, diseño y optimización de enzimas mutadas para procesos biotecnológicos.

1.2. Desarrollo de la Metodología (Pasos a seguir)

Entre las técnicas de regresión usadas, son de destacar las técnicas lineales debido a su sencillez. En ellas se intenta modelar la actividad biológica como una función lineal multivariada de los descriptores moleculares. Por otra parte, en etapas tempranas del descubrimiento molecular así como del estudio del mecanismo de acción de los fármacos, es suficiente tener una respuesta acerca de la probabilidad con que un fármaco tendrá la actividad o mecanismo bajo estudio, sin predecir el valor exacto. Particularmente, en este trabajo se utilizará el Análisis Discriminante Lineal (LDA) ya que posee la cualidad de ser simple permitiendo la clasificación de objetos en grupos predeterminados basándose en múltiples rasgos. En nuestro caso los objetos serán moléculas y los grupos el grado de actividad o un mecanismo de acción determinado. La estrategia general de trabajo en QSAR con LDA se puede dividir en una serie de pasos que son ilustrados gráficamente en la Figura 1[13, 14]:

1. Recopilación de una serie de datos aleatoria, representativa y estratificada de moléculas con la actividad deseada y un grupo control que no posee la actividad bajo estudio.
2. Selección de los descriptores moleculares a utilizar.
3. Cálculo, mediante un programa computacional, de los descriptores moleculares seleccionados.
4. Utilización de los descriptores calculados a las moléculas recopiladas (serie de entrenamiento) para determinar modelos QSAR en un programa de cálculo estadístico.
5. Validación de los modelos QSAR contrastando la actividad predicha a las moléculas recopiladas (serie de predicción) con su actividad experimental.
6. Uso de los modelos encontrados para predecir la actividad a moléculas no ensayadas con anterioridad.

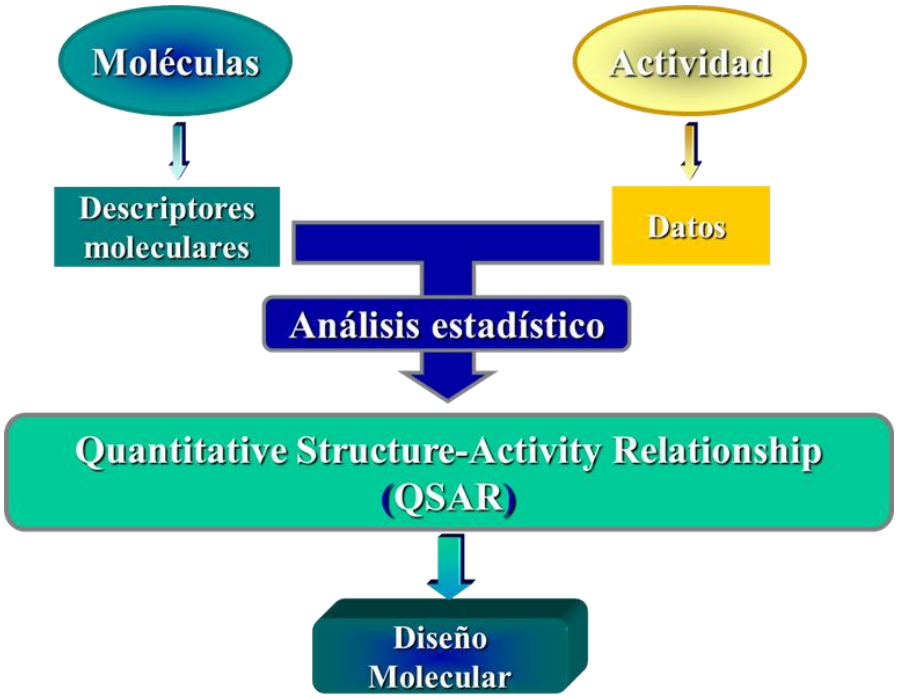


Figura1: Esquema general de trabajo con técnicas QSAR.

2. Los Descriptores Moleculares

El número de moléculas que puede ser obtenido por síntesis orgánica es tan elevado que la probabilidad de seleccionar al azar una molécula que presente la actividad biológica deseada es prácticamente nula. Como consecuencia, ha surgido un gran interés en los métodos teóricos que relacionan la estructura molecular con la actividad biológica, entre ellos especialmente el QSAR. Uno de los pilares para el desarrollo del QSAR lo constituye el proceso de codificar la estructura química mediante descriptores moleculares.[15] Más estrictamente, 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 algún experimento estandarizado. Estas variables pueden ser teóricas o experimentales, pueden describir a la molécula como un todo (descriptores globales) o solo representar un fragmento presente en ella (descriptores fragmentos). Se han definido en la actualidad más de 4000 descriptores moleculares diferentes, véase por ejemplo el programa de cálculo DRAGON y el HANDBOOK de descriptores moleculares recopilados por sus autores (Figura 2).[16]

En la actualidad y en relación con la constante definición de nuevos índices estructurales o descriptores moleculares no se detecta la "explosión" vista en pasados años, observándose cierta tendencia a una aplicación más diversificada e intensiva de los mismos. No obstante, el enorme número de propiedades a estudiar condiciona la continua introducción de nuevos índices basados en otros métodos, con la intención de que los químico-farmacéuticos posean un "arsenal" de descriptores moleculares lo más completo posible.

Este contexto ha propiciado que algunos investigadores se hayan dado a la tarea de crear fórmulas matemáticas de los descriptores moleculares que ofrezcan un cuadro unificado de los mismos, para facilitar su sistematización y estudio.[17-19.] Estas fórmulas pueden, además, indicar la dirección de búsqueda para nuevos descriptores moleculares. En este trabajo se han usado dos programas bien conocidos para el cálculo de descriptores moleculares: el software MARCH INSIDE (MI) y el DRAGON.

Figura 2. Interface del programa DRAGON que calcula más de 4000 descriptores moleculares agrupados en 18 familias diferentes.

2.1. Descriptores del DRAGON

El programa DRAGON ofrece la posibilidad de calcular un gran número de descriptores moleculares agrupados en diferentes familias. A su vez, la lista de descriptores proporcionados puede ser organizada como cerodimensionales (0D), unidimensionales (1D), bidimensionales (2D) y tridimensionales (3D). En nuestro caso y con el objetivo de simplificar la descripción utilizaremos esta última clasificación.[20] Los descriptores calculados en este trabajo son obtenidos con la aplicación de este programa y son cantidades teórico-definidas no habiendo utilizado en ningún caso descriptores experimentales.

Descriptores 0D: describen solamente la constitución de la molécula, pero no aportan información concerniente a la conformación ni al tipo de conectividad presente. Los más simples son, entre otros, el número de átomos de un determinado tipo, el número de enlaces y el peso molecular.

Descriptores 1D: describen fragmentos de las moléculas formados por el agrupamiento de sus átomos constituyentes.

Descriptores 2D: utilizan una función de auto-correlación bidimensional que contiene la topología del grafo y, además, representa la distribución de una propiedad atómica

determinada en la molécula. La propiedad atómica con la que se pesa/pondera el descriptor considera los átomos presentes en la molécula a través de la electronegatividad, la masa atómica, la polarizabilidad atómica, el estado electrotopológico o el volumen de Van der Waals, con lo cual se pueden seleccionar aquellos átomos que proporcionen mayor peso a la variable considerara. Estos descriptores tienen en cuenta las interacciones inter/intramoleculares.

Descriptores 3D: esta clase tiene en cuenta los aspectos conformacionales de la estructura molecular, considerando de esta manera las propiedades estereoquímicas de las moléculas. Para su cálculo se utilizan estructuras moleculares previamente optimizadas con métodos convenientes, tales como el Método de campos de fuerza de la mecánica molecular MM+, en combinación con métodos derivados de la Mecánica cuántica, sean *ab initio* o Métodos de la teoría semiempírica de orbitales moleculares. Entre estos descriptores citamos las cargas atómicas, la energía del orbital molecular más alto ocupado y la energía del orbital molecular más bajo desocupado, entre otros. Un descriptor debe cumplir con un conjunto de características tales como:

i. Fácil cálculo.
ii. Invarianza respecto de la traslación y la rotación.
iii. Invarianza respecto a la numeración de los átomos.
iv. Buena correlación con la propiedad estudiada.
v. Bajo grado de correlación con otros descriptores.

2.2. Descripción del método MARCH-INSIDE.

Cadenas de Markov (CM) es el nombre de una teoría o tipo de modelo matemático definido por Markov.[21] En nuestro trabajo hemos utilizado especialmente el método MARCH-INSIDE (del inglés: Markov Chain Invariants for Network Simulation and Design), el cual emplea las CM para calcular descriptores moleculares mediante una aproximación sencilla a fenómenos tales como [22-27]:

i. distribución de electrones de valencia alrededor de los átomos de una molécula.
ii. propagación de una vibración en una cadena de RNA.
iii. propagación de interacciones electrostáticas superficiales en una proteína viral o en la estructura plegada 3D de una enzima.
iv. paso átomo por átomo de un fármaco desde el plasma a un tejido.
v. interacción paso por paso de un fármaco con su receptor.

3. Estudios Teóricos Fármaco – Proteína

La predicción rápida y precisa de las interacciones entre los fármacos y proteínas es una pieza clave en la combinación de la bioinformática y la investigación del proteoma hacia el descubrimiento de fármacos. Por lo tanto, hay un fuerte incentivo para desarrollar nuevos métodos capaces de detectar estas posibles interacciones fármaco-proteína de manera eficiente.[28] En este sentido, los gráficos y la teoría de redes complejas pueden jugar un papel importante en las diferentes etapas del proceso de modelado con diferentes grados de organización de la materia.[29-36]

3.1. Etapas de los estudios Interacción Fármaco-Proteína.

En una primera etapa, podemos utilizar los gráficos moleculares no sólo para representar y calcular los parámetros estructurales de los fármacos, los ITs, sino también para estimar los parámetros físico-químicos sobre la base de un método gráfico.[37]

En un nivel superior, podemos utilizar los gráficos para representar la estructura de los fármacos-proteínas y calcular los ITs característicos y/o los parámetros físico-químicos de la estructura de las proteínas o las redes de interacciones entre proteínas, véase por ejemplo las obras de Giuliani,[38-44] o las revisiones publicadas en los últimos años.[45] A continuación, se pueden utilizar los ITs y/o los parámetros físico-químicos como entradas para la búsqueda de clasificadores lineales o no lineales capaces de predecir la red -como las estructuras moleculares que presentan o no una propiedad de interés, ver por ejemplo las obras de Caballero y Fernández et al.[46-49]con aplicaciones tanto al campo de los medicamentos y las proteínas o las obras de Zbilut et al.[50,51]

En particular, utilizando los parámetros del fármaco y de la proteína se puede discriminar entre pares fármaco-proteína con alta afinidad (DTPs) y pares fármaco-proteína con nula afinidad (nDTPs). El método QSAR puede convertirse en una herramienta muy útil en este contexto para reducir sustancialmente el tiempo y los recursos que consumen los experimentos.

En una última etapa, la predicción de todos los posibles DTPs/nDTPs de la base de datos forman la red compleja de fármacos y proteínas. Por ejemplo, Yildirim y Goh et al.[52] han construido un grafo bipartito compuesto por la base de datos de la FDA, en

la que están recogidos los fármacos aprobados, y las proteínas unidas por asociaciones binarias. En la red resultante se conectan la mayoría de los fármacos altamente interrelacionados en un componente gigante, con una fuerte agrupación local de fármacos de "tipo similar" de acuerdo con la clasificación de fármacos de tipo Químico-Anatómico-Terapéutica.

El análisis topológico de esta red desde el punto de vista cuantitativo mostró un exceso de "seguimiento de los fármacos", es decir, un exceso de fármacos que se dirigen ya hacía las proteínas específicas.

En esta tesis, presentamos por primera vez dos metodologías nuevas de tipo 2D MI-DRAGON y 3D MI-DRAGON, dos nuevos modelos de interacción fármaco-proteína basados en dos tipos diferentes de software ya conocidos. Utilizamos el software MI para el cálculo de parámetros estructurales en 3D para las proteínas y el software DRAGON para los parámetros de tipo 2D (2D MI-DRAGON) y 3D (3D MI-DRAGON) de todas las proteínas encontradas en el Banco de Drogas (FDA).[53-56]

Estos modelos ofrecen una buena oportunidad para calcular todas las posibles interacciones fármaco-receptor de un fármaco dado y, además, nos permiten volver a construir grandes redes complejas (RC) para estudiar las interacciones fármaco-receptor.

4. Trabajos de Revisión Bibliográfica

En esta sección incluimos 3 trabajos de revisión bibliográfica sobre los antecedentes publicados en la literatura relacionados con los temas objeto de esta tesis. En todos los trabajos revisamos los estudios QSAR con parámetros conceptuales que utilizan análisis de regresión ó LDA, incluyendo estudios teóricos para comprender los requisitos estructurales esenciales para la unión con el receptor. En estas publicaciones de revisión también discutimos modelos 3D y 4D QSAR, CoMFA o CoMSIA con diferentes compuestos. Los estudios realizados se centran en 3 tipos de dianas diferentes con acciones de importancia en el campo de la química médica.

4.1 Revisión de estudios QSAR y bioinformáticos sobre inhibidores de β-*secretasa*

Francisco Javier Prado Prado, Manuel Escobar, Xerardo García Mera.

Current Bioinformatics, **2011**, *6 (1)*, 3-15

El principal factor de riesgo de la enfermedad de Alzheimer (EA) es la edad, afectando a más de la tercera parte de la población mayor de 85 años. Aunque el origen de la EA todavía permanece sin resolver, se ha podido observar en determinado tipo de pacientes, con antecedentes familiares y que desarrollan la enfermedad de un modo temprano, la presencia de proteína precursora de amiloide (APP) en el péptido amiloide β. Esta hipótesis amiloide ha llevado a la realización de números estudios para definir la actividad de la β-secretasa, un enzima con gran relevancia en la EA. Inhibir la acción de dicho enzima es un importante problema cuando se manejan fármacos no derivados de proteínas, por el hecho de que han de atravesar la barrera hematoencefálica. En este sentido, la metodología QSAR podría desempeñar un papel importante en el estudio de estos inhibidores de la β-secretasa. Los modelos QSAR son necesarios para guiar la síntesis de β-secretasa. En el presente trabajo, en primer lugar se revisaron dos servidores como ChEMBL o Protein data bank (PDB) para obtener bases de datos de inhibidores de β-secretasa. A continuación, repasamos trabajos anteriores basados en 2D-QSAR, 3D-QSAR, CoMFA, CoMSIA y técnicas Docking, que estudiaron diferentes compuestos para encontrar los sitios estructurales. Por último, hemos llevado a cabo nuevos estudios QSAR usando métodos de ANNs y el software ModesLab con el fin de comprender los requisitos y sitios estructurales para la unión de los inhibidores con el enzima β-secretasa.

Los estudios teóricos como los modelos QSAR se han convertido en una herramienta muy útil en este contexto para reducir sustancialmente el tiempo y los recursos que consumen los experimentos. Las funciones de la β-secretasa y su implicación en la enfermedad de Alzheimer han dado lugar a una búsqueda activa de potentes y selectivos inhibidores enzimáticos. En este trabajo se puede ver que el desarrollo de nuevos modelos teóricos y QSAR para estudiar este tipo de moléculas es una alternativa diferente a los presentados hasta el momento; la mayoría de estos trabajos presentan estudios de acoplamiento o Docking. Es en este punto donde el QSAR puede desempeñar un papel importante en el estudio de estos inhibidores de la β-secretasa ya que se puede utilizar como herramienta de predicción para el desarrollo de

nuevas moléculas. Por último, hemos desarrollado un nuevo modelo de ANN LNN utilizando los descriptores del software ModesLab, basado en una gran base de datos con cerca de 10.000 medicamentos diferentes obtenidos desde el servidor ChEMBL.

4.2. Revisión de síntesis, ensayos biológicos y estudios teóricos de inhibidores de la β-*secretasa*.

Helena Niño, José Enrique Rodríguez Borges, Xerardo García-Mera y Francisco Javier Prado Prado.

Current Computer Aided Drug Design, **2011**, *7 (4)*, 263-275.

La enfermedad de Alzheimer (EA) es una patología muy compleja, pero con una serie de signos bien caracterizados como la presencia de placas amiloides, lesiones neuropatológicas presentes en el cerebro de pacientes con EA compuestas de péptido β-amiloide. De hecho, una gran cantidad de evidencias sugieren que este péptido es decisivo en la fisiopatología de la EA y es probable que desempeñe un papel en los comienzos de este trastorno neurodegenerativo que a día de hoy todavía permanece como incurable. La enzima BACE-1 es esencial para la generación de β-amiloide. Hoy se sabe que la BACE-1 de ratones knock-out no produce β-amiloide y estos ratones permanecen libres de patologías asociadas a la enfermedad de Alzheimer, incluyendo la pérdida neuronal y ciertos déficits de memoria. El hecho de que BACE-1 inicia la formación de β-amiloide, y la observación de que los niveles de BACE-1 son elevados en la enfermedad son motivos directos y convincentes para desarrollar terapias dirigidas a la inhibición de BACE-1, reduciendo así la concentración de péptido y sus toxicidades asociadas. Los modelos QSAR pueden servir para guiar la síntesis de este enzima.

En este trabajo de revisión nos hemos planteado como objetivo revisar el diseño, síntesis y estudios computacionales diferentes para una serie muy amplia y heterogénea de inhibidores de la β-secretasa. Para ello y en primer lugar realizamos la revisión del diseño, síntesis y análisis biológicos de inhibidores de esta enzima. A continuación, revisamos modelos 2D QSAR, 3D QSAR , CoMFA, CoMSIA y Docking con diferentes compuestos para averiguar los requisitos estructurales de los mismos y finalmente hemos revisado los estudios QSAR realizados utilizando el método de LDA con el fin de comprender las exigencias estructurales esenciales para la unión al receptor de los diferentes inhibidores de la β-secretasa.

En esta publicación hemos desarrollado un nuevo modelo LDA utilizando descriptores de ModesLab, calculados sobre una base de datos de 15.000 fármacos diferentes obtenida del servidor ChemBL.

4.3. Revisión de estudios teóricos y bioinformáticos de inhibidores de la *acetilcolinesterasa*.

Manuel Escobar, Franco Fernández, Xerardo García Mera y Francisco Javier Prado Prado

Current Bioinformatics, **2012**, *in press.*

La enfermedad de Alzheimer (EA) es la enfermedad del siglo XXI, y no sólo es compleja y difícil de tratar sino que hasta ahora es incurable. Tampoco sabemos con certeza que podemos hacer para prevenirla. Es por ello que los tratamientos actuales se centran en múltiples aspectos incluyendo la ayuda a las personas a mantener la función mental, los síntomas conductuales y frenar, retrasar o prevenir la enfermedad. En la actualidad existen cuatro medicamentos aprobados por los EE.UU, Food and Drug Administration (FDA) para el tratamiento de la EA, donepezilo, rivastigmina y galantamina se usan para el tratamiento del Alzheimer en fases de leve a moderada y la memantina se utiliza para tratar los niveles moderado a severo. Estos medicamentos actúan regulando los neurotransmisores (las sustancias químicas que transmiten mensajes entre las neuronas). El tratamiento de la EA por los precursores de acetilcolina y los agonistas colinérgicos se ha mostrado ineficaz e incluso ha originados graves efectos secundarios. La hidrólisis de la ACh por la AChE causa la terminación de la neurotransmisión colinérgica. Por lo tanto, los compuestos que inhiban significativamente la AChE podrían aumentar los niveles de ACh y paliar la EA. Sin embargo, estos medicamentos no cambian el proceso de la enfermedad y pueden ayudar unos pocos meses o incluso unos pocos años.

En este sentido, la metodología QSAR podría desempeñar un papel importante en el estudio de estos inhibidores de la acetilcolinesterasa (ACE). Los modelos QSAR pueden contribuir a guiar la síntesis de la AChE. En este trabajo hemos revisado diferentes estudios de bioinformática y estudios teóricos de inhibidores de la AChE, estudios de diseño y cálculo de una serie muy grande y heterogénea de los inhibidores de este enzima. En primer lugar, revisamos los métodos 2D QSAR, 3D QSAR, CoMFA, CoMSIA y de docking con diferentes compuestos para averiguar los requisitos estructurales. A continuación, revisamos los estudios QSAR utilizando el método de LDA con el fin de comprender las exigencias estructurales esenciales para la unión con el receptor de dichos inhibidores de la AChE, utilizando descriptores de ModesLab de una base de datos de 10.000 fármacos diferentes del servidor ChemBL.

5. *Objetivos*

5.1. Objetivos Generales:

I. Desarrollar nuevos modelos QSAR aplicables a la predicción de la actividad biológica de compuestos contra una única diana o múltiples dianas, modelos QSAR multi-target (mt-QSAR), de interés en química farmacéutica, microbiología, y parasitología.

II. Desarrollar nuevas metodologías haciendo uso de varios programas de cálculo de descriptores moleculares para la construcción de Redes Complejas de compuestos útiles en estudios de Bioinformática a partir de modelos QSAR ó mt-QSAR.

5.2. Objetivos específicos:

I. Desarrollar modelos QSAR para la predicción de inhibidores de la β-*secretasa.*

II. Desarrollar modelos QSAR para la predicción de inhibidores de la *acetilcolinestarasa.*

III. Desarrollar modelos mt-QSAR para la predicción de inhibidores de la *acetilcolinestarasa*.

IV. Desarrollar nuevas metodologías para generar nuevos modelos que permitan estudiar las interacciones fármaco-proteína.

V. Desarrollar una metodología QSAR/QSPR usando nuevos índices topológicos para construir redes de interacción fármaco-proteína.

VI. Desarrollar un nuevo método para cuantificar numéricamente la calidad de las conexiones entre vértices basándose en los índices Markov para redes fármaco-proteína.

6. Resultados y Discusión

En este punto se presentarán todos los resultados obtenidos en forma de artículos de investigación ya publicados por el autor. Los 4 artículos presentados están agrupados de acuerdo al objetivo específico que cumplimentan. Para cada artículo se presenta una breve sección explicativa en Español de su importancia y los resultados alcanzados. En el apartado "9. Anexos (Publicaciones)" de esta Tesis se adjuntan las publicaciones correspondientes en el idioma en que fueron publicadas.

6.1. 2D MI-DRAGON: un nuevo modelo para estudiar las interacciones proteína-ligando y el estudio teórico-experimental de la red fármaco-proteína obtenida de la base de datos de la FDA de EE.UU., estudio de oxoisoaporfinas como inhibidores de la *MAO-A* y proteínas de parásitos en humanos.

Francisco Prado-Prado, Xerardo García-Mera, Manuel Escobar, Eduardo Sobarzo-Sánchez, Matilde Yañez, Pablo Riera-Fernández and Humberto González-Díaz.
European Journal of Medicinal Chemistry, **2011**, *46 (12)*, 5838-5851.

Hay muchos posibles pares de fármacos-proteínas que pueden tener lugar o no (DTP/nDTPs) entre las drogas con alta afinidad/no afinidad con proteínas diferentes. Por este motivo resulta costoso en términos de tiempo y recursos, por ejemplo, la determinación de todas las posibles interacciones entre ligandos de proteínas para un solo medicamento. En este aspecto, podemos utilizar el QSAR para llevar a cabo la predicción de DTP. Desafortunadamente, casi todos los modelos QSAR predicen la actividad contra un solo objetivo, proteína o diana. Para solucionar este problema se puede desarrollar un multi-target QSAR (mt-QSAR). En este trabajo, presentamos la técnica 2D MI-DRAGON un nuevo predictor de DTP basado en dos programas de software diferentes conocidos. Utilizamos el software MI para el cálculo de parámetros estructurales en 3D para las proteínas y el software DRAGON, uno de los más completos y basado en cálculos con más de 1600 parámetros, para calcular dichos descriptores moleculares en 2D mostrando todos los fármacos que tienen interacción conocida con proteínas presentes en la base de datos de la FDA. Un modo de desarrollar este tipo de ensayos en mt-QSAR consiste en incorporar en las ecuaciones QSAR parámetros de la estructura de las dianas (proteínas, DNA, RNA, etc..) añadidos a los parámetros estructurales de los fármacos presentes en el clásico QSAR. Ambas clases de parámetros fueron utilizados como inputs en diferentes algoritmos de redes neuronales artificiales para buscar un modelo no lineal mt-QSAR muy preciso. El mejor modelo ANN encontrado es un perceptrón multicapa (MLP) cuyo perfil es MLP 21:21-31-1:1, el cual clasifica correctamente 303 de 339 DTP (sensibilidad = 89,38%) y 480 de 510 nDTPs (especificidad = 94,12%), que corresponde al promedio de la serie de entrenamiento = 92,23%. La validación del modelo se llevó a cabo por medio de una serie externa de predicción con una sensibilidad = 92,18% (625/678 DTP) y con una especificidad = 90,12% (730/780 nDTPs) y un promedio = 91,06%.

El modelo 2D MI-DRAGON ofrece una buena oportunidad para calcular de forma rápida todos los DTP posibles de un fármaco lo que nos permite reconstruir grandes redes complejas de DTP. Por ejemplo, hemos reconstruido con la base de datos de la FDA una red compleja con 855 nodos de 519 medicamentos + 336 objetivos. Hemos predicho una red compleja con topología similar (valores observados y previstos de media distancia es igual a 6,7 frente a 6,6. Esta red compleja puede ser utilizada para explorar grandes bases de datos de DTP con el fin de descubrir los nuevos medicamentos y / o proteínas.

Finalmente, se ilustra con un estudio teórico-experimental el uso práctico del modelo 2D MI-DRAGON. En este trabajo se realiza la predicción, síntesis y evaluación farmacológica de 10 oxoisoaporfiinas diferentes con actividad inhibitoria de la MAO-A. El compuesto más activo OXO_5 presentó una IC_{50} = 0,83 μM, notablemente mejor que la clorgilina que se tomó como fármaco control.

Es posible buscar excelentes predictores de DTP utilizando como entrada parámetros estructurales de los fármacos y proteínas calculados con diferentes programas y en combinación con modelos ANN. El modelo 2D MI-DRAGON, basado en los parámetros estructurales de los fármacos calculada con el DRAGON y los parámetros de proteínas calculado con el MI, predice correctamente los DTPs de 500 fármacos diferentes aprobados por la FDA con una precisión mayor del 90%.

El modelo 2D MI-DRAGON también es útil para construir y desarrollar nuevas redes complejas computacionalmente construidas, las cuales ofrecen una alternativa para descubrir nuevos medicamentos y proteínas y explorar la selectividad y la toxicidad de los fármacos.

En este trabajo estas conclusiones se ejemplifican a través del estudio experimental y teórico de las nuevas isoaporfinas que presentan actividad inhibitoria de la *MAO-A*.

6.2. 3D MI-DRAGON: nuevo modelo para la reconstrucción de la red fármaco-proteína de la base de datos de la FDA y estudios teóricos experimentales de los derivados de la rasagilina como inhibidores de la *acetilcolinesterasa*.

Francisco Prado-Prado, Xerardo García-Mera, Manuel Escobar, Nerea Alonso, Olga Caamaño,[1] Matilde Yañez and Humberto González-Díaz.

European Journal of Organic Chemistry, **2012**. *Submitted*

Las enfermedades neurodegenerativas se han incrementado de manera notable en los últimos años. Muchos de los fármacos que se utilizan en el tratamiento de dichas enfermedades presentan características específicas estructurales en 3D. Una proteína importante en este sentido es la *acetilcolinesterasa*, que es la diana de muchos fármacos usados en la Enfermedad de Alzheimer. En consecuencia, la predicción de las interacciones fármaco-proteína (DTP/nDTPs) entre nuevos fármacos candidatos con estructura 3D específica y proteínas adquiere gran relevancia. Por ello, podemos utilizar técnicas mt-QSAR para realizar la predicción de DTPs. Desafortunadamente, muchos de los modelos QSAR previamente desarrollados para predecir DTPs toman en consideración únicamente la información estructural 2D y codifican la actividad frente a una sola diana o proteína. Para contribuir a la resolución de este problema, hemos desarrollado un modelo 3D multi-target QSAR (3D mt-QSAR). En este trabajo, se introduce la técnica de 3D MI-DRAGON un nuevo modelo que predice los DTPs basándose en la utilización de dos clases de software conocidos. Así utilizamos el software MI y DRAGON 3D para el cálculo de los parámetros estructurales de los fármacos y las proteínas, respectivamente. Ambas clases de parámetros 3D se utilizan como materia prima para entrenar la ANN, utilizando los datos como referencia para construir la red compleja formada por todos los DTPs encontrados como fármacos-proteínas que han sido aprobadas por la FDA. El conjunto de datos se ha descargado del Drug Bank. El mejor modelo 3D mt-QSAR encontrado es un ANN de tipo Multi-Layer Perceptron (MLP) con el perfil MLP 37:37-24-1:1, el cual clasifica correctamente 274 de 321 DTP (sensibilidad = 85,35%) y 1041 de 1190 nDTPs (especificidad = 87,48%), que corresponde al promedio = 87,03%. La validación del modelo se ha llevado a cabo con una serie de predicción externa con sensibilidad = 84,16% (542/644 DTP). Especificidad = 87,51% (2039/2330 nDTPs) y Promedio = 86,78%. Las nuevas redes complejas de los DTPs han sido reconstruidas a partir de la base de datos de la FDA y

pueden ser utilizadas para explorar grandes bases de datos de DTPs con el fin de descubrir nuevos fármacos y/o proteínas.

Se llevó a cabo un estudio teórico-experimental que ilustra el uso práctico del modelo 3D MI-DRAGON. En primer lugar, describimos la predicción y evaluación farmacológica de una nueva serie de 22 derivados de rasagilina con potencial actividad inhibitoria de *acetilcolinesterasa*.

El modelo 3D MI-DRAGON está basado en los parámetros estructurales de los fármacos calculados con el DRAGON y los parámetros de proteínas calculado con el MI. Es posible buscar excelentes modelos que predigan DTPs utilizando como entrada parámetros estructurales de los fármacos y proteínas calculados con diferentes programas y en combinación con modelos de ANN diferentes. La combinación de MI y de DRAGON genera una ANN que hace posible lograr un método mt-QSAR que permite predecir con un porcentaje de clasificación superior al 85% la probabilidad de que los fármacos se unan a más de 500 proteínas conocidas o dianas terapéuticas ya publicadas y aprobadas por la FDA.

El modelo 3D MI-DRAGON también es útil para montar redes complejas de DTPs desarrolladas computacionalmente, las cuales son una alternativa para descubrir nuevos medicamentos o nuevas dianas terapéuticas, y explorar una mayor selectividad de los fármacos cara a esas dianas.

6.3. Nuevos modelos Markov-Entropía de Shannon para evaluar la calidad de la conectividad en redes complejas: De las redes moleculares a las rutas metabólicas, redes parásito-hospedador, redes neurales, redes industriales y redes sociales-jurídicas.

P. Riera-Fernández, C. R. Munteanu, M. Escobar, F. Prado-Prado, R. Martín-Romalde, D. Pereira, K. Villalba, A. Duardo-Sánchez and H. González-Díaz.

Journal of Theoretical Biology, **2012**, *293*, 174-188.

Existen muchos métodos teóricos y experimentales para determinar las conexiones entre vértices a la hora de construir redes complejas. Lamentablemente, la reevaluación de la red entera mediante métodos experimentales es cara en términos de tiempo y recursos. Por tanto, resulta muy interesante desarrollar métodos computacionales que permitan evaluar la calidad de las conexiones (la probabilidad de que la conexión sea cierta o no) *a posteriori* de la construcción de la red. En este trabajo se desarrolla un nuevo método para cuantificar numéricamente la calidad de las conexiones entre vértices basándose en los índices de Markov – Entropía de Shannon de orden k (θ_k) para los nodos de la red.

El método puede resumirse como sigue:

a) Se calculan los valores θ_k para cada uno de los nodos de una red ya construida.
b) Se usa un análisis lineal discriminante para buscar una ecuación que discrimine entre pares conectados (observados experimentalmente) y pares no conectados.
c) El nuevo modelo se valida mediante una serie de datos externa.
d) La ecuación obtenida se usa para reevaluar la conectividad de la red, conectando o desconectando nodos en función de los valores obtenidos.

Los modelos obtenidos usados para evaluar distintos tipos de redes produjeron los siguientes resultados con un porcentaje de precisión media mayor al 72%.

Se desarrolló un modelo para evaluar la calidad de las conexiones en la red de fármacos-receptor construida a partir de una lista de fármacos aprobados por la FDA. En este último caso los valores de θ_k fueron calculados para tres tipos de redes moleculares que representan distintos tipos de organización: red de la estructura de los fármacos (conexiones átomo-átomo), redes de la estructura de los receptores proteicos (conexiones entre aminoácidos) y red fármaco-receptor (conexiones entre los fármacos y sus receptores proteicos). La precisión media de este modelo fue de 76,3 %.

6.4. De los modelos QSAR de fármacos a redes complejas: Revisión del estado del arte e introducción de nuevos índices de Markov-Momentos espectrales.

P. Riera-Fernández, R. Martín-Romalde, F. Prado-Prado, M. Escobar, C. R. Munteanu, R. Concu, A. Duardo-Sánchez and H. González-Díaz.

Current Topics in Medicinal Chemistry, **2012**, *12 (8),* 927-960.

Las aplicaciones de los modelos QSAR/QSPR se han restringido tradicionalmente al estudio de pequeñas moléculas. Por otra parte, el estudio de moléculas de gran tamaño (como el DNA o proteínas) y de otros sistemas complejos ha llevado a la construcción de redes complejas de gran tamaño. Una interesante pregunta a responder es si el uso de los índices topológicos (números que codifican la información estructural del grafo molecular) puede extenderse a grafos de gran tamaño que representan sistemas tan diversos como redes fármaco-receptor, redes parásito-hospedador, redes de propagación de enfermedades, redes metabólicas, redes cerebrales o redes sociales-jurídicas y si se pueden construir modelos QSAR/QSPR basados en dichos índices para predecir la conectividad de las redes. En el presente trabajo se realiza una revisión de algunas de las bases de datos, software, modelos QSAR/QSPR y redes complejas relacionadas con estudios de fármacos o sus receptores (proteínas, parásitos…). Además, se revisan índices topológicos que han sido utilizados para describir la estructura de fármacos o de redes complejas de gran tamaño y se define un nuevo tipo de índices topológicos basados en el uso de cadenas de Markov, los momentos espectrales estocásticos. Por último, se desarrollan modelos QSAR/QSPR para diferentes clases de redes usando estos últimos índices topológicos. El porcentaje de buena clasificación de los modelos obtenidos para las distintas redes estudiadas son: red fármaco-receptor (77,8 %), redes parásito-hospedador (88,7 %), red de propagación de la fasciolosis (90,6 %), red del metabolismo de *Caenorhabditis elegans* (96,3 %), red del cerebro de macaco (94,8 %) y red de las leyes financieras españolas (90,4 %).

7. Conclusiones

7.1. Conclusión general:

Exponemos las conclusiones específicas, en correspondencia con los objetivos trazados, diferenciadas en dos grupos, en función de la naturaleza de los estudios realizados: 1) estudios QSAR/mt-QSAR de inhibidores de enzimas, y 2) estudio mt-QSAR y desarrollo de nuevas redes complejas de interacciones fármaco-proteína.

7.2. Conclusiones específicas:

I. Hemos desarrollado dos modelos QSAR para la predicción de inhibidores de la β-*secretasa*.

II. Hemos desarrollado un modelo QSAR para la predicción de inhibidores de la *acetilcolinestarasa*.

III. Hemos desarrollado un modelo mt-QSAR para la predicción de inhibidores de la *acetilcolinestarasa*.

IV. Hemos desarrollado dos nuevas metodologías que permiten estudiar las interacciones fármaco-proteína, el modelo 2D MI-DRAGON y el modelo 3D MI-DRAGON, las cuales presentaron valores superiores al 85% en precisión media. A partir de estos modelos se han obtenido nuevas redes complejas sobre la interacción fármaco-receptor, enfocándonos al diseño de nuevos fármacos con actividad neuroprotectora.

V. Hemos desarrollado un nuevo método para cuantificar numéricamente la calidad de las conexiones entre vértices basándose en los índices Markov para redes fármaco-proteína con un 76.3% de precisión media.

VI. Hemos desarrollado usando nuevos índices topológicos para construir redes de interacción fármaco-proteína una metodología QSAR/QSPR cuyo porcentaje de buena clasificación es igual a 77.8%.

8. Referencias Bibliográficas

1. Olson, R.E. *Annu. Rep. Med. Chem.* **2000**, *35*, 31-40.
2. Woodgett, J.R. *EMBO J.* **1990**, *9 (8)*, 2431-2438.
3. Troussard, A.A; Tan, C., Yoganathan, T.N. *Dedhar S. Molecular Cell Biology*. **1999**, *19*, 7420-7427.
4. Turenne, G.A.; Price, B.D. *BMC Cell. Biol*. **2001**, *2*, 12-21.
5. Hoeflich, K.P.; Luo, J.; Rubie, E.A.; Tasao, M.S.; Jin, O.; Woodgett, J.R. *Nature*. **2000**, *406*, 86-90.
6. Hooper, C.; Killick, R.; Lovestone, S. *J. Neurochem*. **2008**, *104*, 1433-1439.
7. MacAulay, K.; Doble, B.W.; Patel, S.; Hansotia, T.; Sinclair, E.M.; Drucker, D.J. et al. *Cell Metab*. **2007**, *6*, 329-337.
8. Droucheau, E.; Primot, A.; Thomas, V.; Mattei, D.; Knockaert, M.; Richardson, C.; et al. *Biochim Biophys. Acta*. **2004**, *1697*, 181-196.
9. Kubinyi, H.; Taylor, J.; Ramdsen, C. *Quantitative Drug Design, in Comprehensive Medicinal Chemistry*. Ed. C. Hansch. Pergamon. **1990**, *4*, 589-643.
10. Lutz, M.W.; Menius, J. A.; Laskody, R.G.; Domanico, P.L.; Goetz, A.G.; Saussy, D. L.; Rimele, T. *Network Science*. **1996**, *2*, Issue 9, September.
11. Loew, G.H.; Villar, H.O.; Alkorta, Y. *Pharm. Res*. **1993**, *10*, 475-486.
12. Wess, G. *Drug Discovery Today*. **1996**, *1*, 529-535.
13. Kier, L.B.; Hall, L.H. *Topological Indices and Related Descriptors in QSAR and QSPR. Gordon and Breach, Amsterdam*, **1999**.
14. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors.* Mannhold, R.; Kubinyi, H.; Timmermann, H. Ed., Wiley-VCH: Weinheim, **2000**.
15. Devillers, J.; Balaban, A.T. *Topological Indices and Related Descriptors in QSAR and Drug Design*, Amsterdan, **2000**, pp 3-41.

16. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*. Wiley VCH, Weinheim, Germany. **2000**.
17. Estrada, E.; *Chem. Phys. Lett* .**2001**, *336*, 9890-9895.
18. Kier, L.B.; Hall, L.H. *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach Sci. Pub.: Amsterdam, **1999**, pp 455-489.
19. Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17-20.
20. Helguera, A.M.; Combes, R.D.; González, M.P.; Cordeiro, M.N.D.S.. *Curr. Top. Med. Chem*. **2008**, *8 (18)*, 1628-1655.
21. Markov, A.A. *Bull. Soc. Phys. Math. Kasan* **1906**, *15*, 155-165.
22. Bharucha-Reid, A.T. *Elements of Theory of Markov Process on the Application, McGraw-Hill Series in Probability and Statistic, McGraw-Hill Book Company*, New York. **1960**, 167-434.
23. Freund, J.A.; Poschel, T. *Eds. Stochastic Processes in Physics, Chemistry, and Biology. In: Lect. Notes Phys. Springer-Verlag*, Berlin, Germany **2000**.
24. González-Díaz, H.; Prado-Prado, F.; Ubeira, F.M. *Curr. Top. Med. Chem*. **2008**, *8 (18)*, 1676-90.
25. González-Díaz, H.; Duardo-Sanchez, A.; Ubeira, F.M.; Prado-Prado, F.; Pérez-Montoto, L.G.; Concu, R.; Podda, G.; Shen, B. *Curr. Drug. Metab.* **2010** *May; 11 (4)*, 379-406.
26. González-Díaz, H.; González-Díaz, Y.; Santana, L.; Ubeira, F.M.; Uriarte, E. *Proteomics*. **2008**, *Feb, 8 (4)*, 750-778.
27. González-Díaz, H.; Vilar, S; Santana, L.; Uriarte, E. *Curr. Top. Med. Chem.* **2007**, *7 (10)*, 1015-1029.
28. Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M., *Bioinformatics* **2008**, *24 (13)*, i232-240.
29. Giuliani, A. *BMC Genomics*. **2010**, *11* Suppl 1, S2.
30. Dhar, P. K.; Giuliani, A. *Sys. Synth. Biol.* **2010**, *4, (1)*, 7-13.
31. Bornholdt, S.; Schuster, H. G. *Handbook of Graphs and Complex Networks: From the Genome to the Internet. WILEY-VCH GmbH & CO. KGa.* Wheinheim, **2003**.
32. Estrada, E. *Proteomics* **2006**, *6 (1)*, 35-40.
33. Estrada, E. *J. Proteome Res.* **2006***, 5 (9)*, 2177-2184.
34. Réka, A.; Barabasi, A.L. *Rev. Mod. Phys*. **2002**, *74 (1)*, 47-97.
35. Barabasi, A. L.; Oltvai, Z. N. *Nat. Rev. Genet*. **2004**, *5 (2)*, 101-113.

36. Barabasi, A. L. *N. Engl. J. Med*. **2007**, *357 (4)*, 404-407.
37. González-Díaz, H.; Vilar, S.; Santana, L.; Uriarte, E. *Curr. Top. Med. Chem*. **2007**, *7 (10)*, 1025-1039.
38. Giuliani, A.; Di Paola, L.; Setola, R. *Curr. Proteomics* **2009**, *6 (4),* 235-245.
39. Krishnan, A.; Zbilut, J. P.; Tomita, M.; Giuliani, A. *Curr. Protein Pept. Sci.* **2008**, *9 (1)*, 28-38.
40. Krishnan, A.; Giuliani, A.; Zbilut, J. P.; Tomita, M. *PLoS ONE*. **2008**, *3 (5)*, e2149.
41. Palumbo, M. C.; Colosimo, A.; Giuliani, A.; Farina, L. *FEBS Lett*. **2007**, *581 (13)*, 2485-2489.
42. Krishnan, A.; Giuliani, A.; Zbilut, J. P.; Tomita, M. *J. Proteome Res.* **2007**, *6 (10)*, 3924-3934.
43. Krishnan, A.; Giuliani, A.; Tomita, M. *PLoS ONE*. **2007**, *2 (6*), e562.
44. Tun, K.; Dhar, P. K.; Palumbo, M. C.; Giuliani, A. *BMC Bioinformatics.* **2006**, 7-24.
45. Vilar, S.; González-Díaz, H.; Santana, L.; Uriarte, E. *J. Theor. Biol*. **2009**, *261 (3)*, 449-458.
46. Caballero, J.; Fernández, M. *Curr. Top. Med. Chem*. **2008**, *8 (18)*, 1580-1605.
47. Fernández, L.; Caballero, J.; Abreu, J. I.; Fernández, M. *Proteins* **2007**, *67*, 834–852.
48. Fernández, M.; Caballero, J.; Fernández, L.; Abreu, J. I. *J. Mol. Graph. Model*. **2007**, *26 (4),* 748-759.
49. Fernández, M.; Caballero, F.; Fernández, L.; Abreu, J. I.; Acosta, G. *Proteins* **2008**, *70 (1*), 167-175.
50. Zbilut, J. P.; Giuliani, A.; Colosimo, A.; Mitchell, J. C.; Colafranceschi, M.; Marwan, N.; Webber, C. L., Jr.; Uversky, V. N. *J. Proteome Res.* **2004**, *3 (6*), 1243-1253.
51. Zbilut, J. P.; Colosimo, A.; Conti, F.; Colafranceschi, M.; Manetti, C.; Valerio, M.; Webber, C. L., Jr.; Giuliani, A. *Biophys. J.* **2003**, *85 (6)*, 3544-3557.
52. Yildirim, M. A.; Goh, K. I.; Cusick, M. E.; Barabasi, A. L.; Vidal, M. *Nat. Biotechnol*. **2007**, *25 (10)*, 1119-1126.
53. Kirchmair, J.; Markt, P.; Distinto, S.; Schuster, D.; Spitzer, G.M.; Liedl, K.R.; Langer, T.; Wolber, G. *J. Med. Chem*. **2008**, *51*, 7021-7040.

54. Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A.C.; Wishart, D.S. *Nucleic Acids Res*. **2011**, *39*, 1035-1041.
55. Wishart, D.S.; Knox, C.; Guo, A.C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. *Nucleic Acids Res*. **2008**, *36*, 901-906.
56. Wishart, D.S.; Knox, C.; Guo, A.C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. *Nucleic Acids Res*. **2006**, *34*, 668-672.

9. Anexos (Publicaciones)

A continuación se presentan las diferentes publicaciones que se recogen en la Tesis siguiendo el orden de presentación de las mismas establecido en la memoria.

Review of Bioinformatics and QSAR Studies of β-Secretase Inhibitors

Francisco Prado-Prado*, Manuel Escobar-Cubiella and Xerardo García-Mera

Department of Organic Chemistry, University of Santiago de Compostela, Spain

Abstract: Alzheimer disease (ADa) is the most common form of senile dementia, and it is characterized pathologically by decreased brain mass. An important problem to inhibiting β-secretase, is to cross the blood-brain barrier (BBB) using drugs not derived from proteins and thus more efficient to design drugs to treat Alzheimer's disease. In this sense, quantitative structure-activity relationships (QSAR) could play an important role in studying these *β*-secretase inhibitors. QSAR models are necessary in order to guide the *β*-secretase synthesis. In the present work, we firstly revised two servers like ChEMBL or PDB to obtain databases of *β*-secretase inhibitors. Next, we review previous works based on 2D-QSAR, 3D-QSAR, CoMFA, CoMSIA and Docking techniques, which studied different compounds to find out the structural requirements. Last, we carried out new QSAR studies using Artificial Neural Network (ANN) method and the software ModesLab in order to understand the essential structural requirement for binding with receptor for *β*-secretase inhibitors.

Keywords: QSAR, CoMSIA, COMFA, docking, topological indices, *β*-secretase inhibitors, alzheimer's disease (AD).

1. INTRODUCTION

Alzheimer disease (ADa) is the most common form of senile dementia, and it is characterized pathologically by decreased brain mass, extracellular senile plaques, and intracellular neurofibrillary tingles [1]. The major risk factor for AD is age, and it affects more than a third of all people that reach 85 years of age. Although the pathogenesis of AD is still controversial, rare human mutations that lead to familial early onset AD are found in genes that lead to increased expression or processing of the amyloid precursor protein (APP) into amyloid β peptide (Aβ), which is the major component of senile plaques [2-4]. The so-called "amyloid hypothesis" for AD pathogenesis is supported by transgenic mouse models that overexpress mutant APP and genes involved in APP processing, which lead to the production of senile plaques and cognitive impairment [5-8] (see Fig. **1**).

Prior to its identification, numerous studies were undertaken to define the characteristics of β-secretase activity. Although the majority of body tissues exhibit β-secretase activity [9], highest activity levels were observed in neural tissue and neuronal cell lines [10]. β-Secretase also called BACE1 (β-site of APP cleaving enzyme) (see Fig. **2**), is an aspartic-acid protease important in the pathogenesis of Alzheimer's disease, and in the formation of myelin sheaths in peripheral nerve cells [5-8]. The main problem inhibiting β-secretase, is to cross the blood-brain barrier (BBB) using drugs not derived from proteins and thus more efficient to design drugs to treat Alzheimer's disease. There are two barriers in the brain: the blood-cerebrospinal fluid barrier and the BBB. The BBB consists of the endothelial cells that form the brain capillaries. It is restrictive to penetration of molecules, owing to tight junctions, lack of fenestrations, negative surface polarity, and the high level of efflux transporters [11].

In this part, Chemoinformatics and Bioinformatics methods may play an important role in the study of BACE1 inhibitors, Quantitative Structure-Activity Relationships (QSAR) studies are used as predictive tools for the molecular development [12, 13]. Up to today, there are near 1600 molecular descriptors that, in principle, can be generalized and used to solve the former problem [14]. Many of these indices are known as molecular Topological Indices (TIs) or simply invariants of a molecular graph. Unfortunately, QSAR studies are generally based on databases considering only structurally parent compounds acting against one single microbial species. In a recent review, our group have discussed recent advances in the field [15]. In addition to QSAR, Bioinformatics and Chemoinformatics methods useful to study β-secretase may include techniques like Comparative Molecular Field Analysis (CoMFA), drug-target Docking, Sequence Alignment (SA) or other methods. In a recent, preliminary review in the field published in Proteomics in 2008 was discussed the use of these methods but only from the point of view of proteins [16]. Almost all QSAR techniques are based on the use of molecular descriptors, which are numerical series that codify useful chemical information and enable correlations between statistical and biological properties [17, 18]. On the other hand, QSAR models can be used to explore the relationships between the structural spaces of compounds as inhibitors for specific enzymes, such as MAO inhibitors [19], HIV-1 integrase inhibitors [20], and/or protease inhibitors [21] or tyrosinase inhibitors [22-24]. In fact, recently, the field has moved from small molecules to proteins and other systems. See for instance, recent review issues in Current Topics in Medicinal Chemistry [25-34], Current Proteomics [35-41], Current Drug Metabolism [42-50] and Current Pharmaceutical Design [51-60]. In the present work, we firstly revised servers like ChEMBL (http://www.ebi.ac.uk/ChEMBLdb/) or PDB (http://www.-

*Address correspondence to this autor at the Department of Organic Chemistry, University of Santiago de Compostela, Spain; Tel: + 881814940; Fax: 34-981-594912; E-mail: francisco.prado@usc.es.

1574-8936/11 $58.00+.00

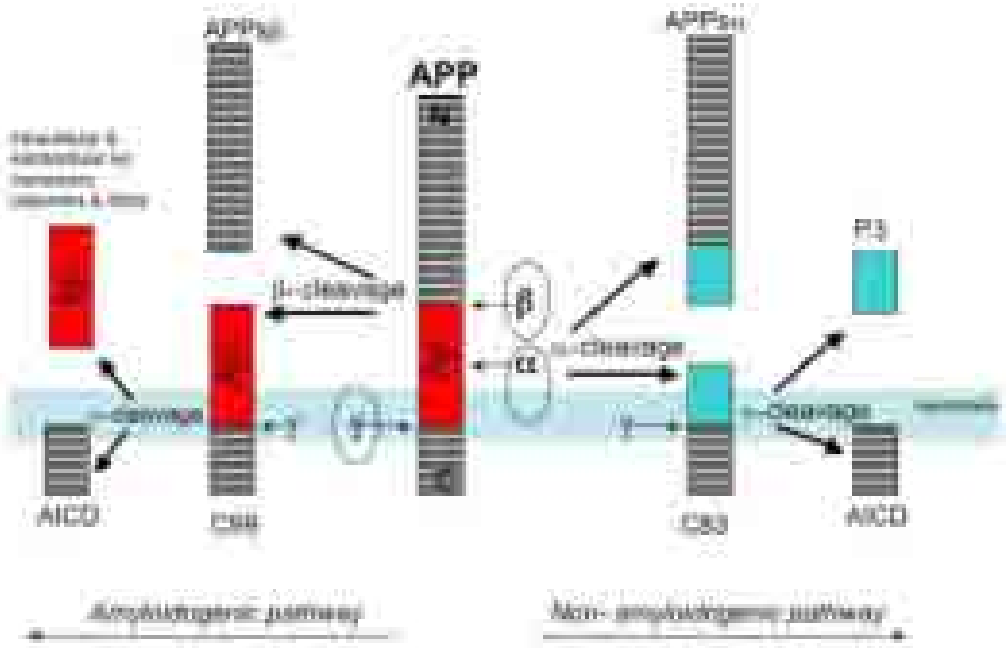


Fig. (1). APP metabolism by the secretase enzymes.

Fig. (2). 3D of β-secretase.

pdb.org/pdb/home/home.do) to obtain databases of *β*-secretase inhibitors. Next, we review previous works based on 2D-QSAR, 3D-QSAR, CoMFA, CoMSIA and Docking techniques, which studied different compounds to find out the structural requirements. Last, we carried out new QSAR studies using Artificial Neural Network (ANN) method and the software ModesLab [61] in order to understand the essential structural requirement for binding with receptor for *β*-secretase inhibitors.

2. THEORETICAL STUDIES FOR β-SECRETASE INHIBITORS

In this section we updated the contents presented in our recent review published in Current Drugs Metabolim [49]. The high number of possible candidates to β-secretase inhibitors creates the necessity of Quantitative Structure-Activity Relationship models in order to guide the β-secretase inhibitor synthesis. In this work, we revised diffe-

rent computational studies for a very large and heterogeneous series of β-secretase. First, we revised databases for bioinformatics studies of secretase for QSAR studies with conceptual parameters. Next, using method of regression analysis; and QSAR studies in order to understand the essential structural requirement for binding with receptor. Next, we review 3D QSAR, CoMFA, CoMSIA and Docking with different compound to find out the structural requirements for β-secretase inhibitors.

2.1. Databases for Bioinformatics Studies of Secretase

2.1.1. ChEMBL Dataserver of β-Seceretase Inhibitors for QSAR Studies

ChEMBL [62] is a database of bioactive drug-like small molecules, it contains 2-D structures, calculated properties (e.g. logP, Molecular Weight, Lipinski Parameters, etc.) and abstracted bioactivities (e.g. binding constants, pharmacology and ADMET data) (see Fig. **3**). In this review, the search for β-secretase inhibitors by ChEMBL is important for the development of new theoretical studies and QSAR models or docking studies. The β-secretase database attempt to normalise the bioactivities into a uniform set of end-points and units where possible, and also to tag the links between a molecular target and a published assay with a set of varying confidence levels. The data is abstracted and curated from the primary scientific literature, and cover a significant fraction of the SAR and discovery of modern drugs. Additional data on clinical progress of compounds is being integrated into ChEMBL at the current time. ChEMBL search by:

- Search target data via keyword, protein sequence search (BLAST), or by navigating the target classification hierarchy.
- Search compound data with lists of keywords, SMILES strings, or compound identifiers. Substructure and similarity searching functionality is available also.
- Search assay data via keyword search using the main search bar.

The ChEMBL database (ChEMBLdb) contains medicinal chemistry bioassay data, integrated from a wide variety of sources (the literature, deposited data sets, other bioassay databases). Subsets of ChEMBLdb, relating to particular target classes, or disease areas, are exported to smaller databases like β-seceretase inhibitors. ChEMBL target search results for secretase enzymes was 11 hits, see Table **1**. The table shows the target ID, the description of the target, the organism, the compounds and end points. In the case of the Beta-secretase 1, presents 2 157 compounds, and 3 360 endpoints. These separate data sets, (see Fig. **4**)**,** and the entire ChEMBLdb, are available either via ftp downloads or .xls files download, or via bespoke query interfaces, tailored to the requirements of the scientific communities with a specific interest in these research areas**.**

2.1.2. PDB Structures of β-Secretases for DOCKING Studies

The Protein Data Bank (PDB) www.pdb.org archive is the single worldwide repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids, (see Fig. **5**). These are the molecules of life that are found in all organisms including bacteria, yeast, plants, flies, other animals, and humans. Understanding the shape of a molecule helps to understand how it works. This knowledge can be used to help deduce a structure's role in human health and disease, and in drug development. The structures in the archive range from tiny proteins and bits of DNA to complex molecular machines like the ribosome. 3D

Fig. (3). ChEMBL database server.

Fig. (4). ChEMBL bioactivities results for β-secretase.

structures are important for the development of Docking. In this review we report the use of pdb as a key to developing drug-protein interaction studies and docking of the β-secretase inhibitors.

First, we seek in the PDB, secretase and got 157 hits or secretase-related proteins; to obtain a more thorough search, search Homologue Removal realized with a - 30% Cutoff Identity and getting only eight hits. Each peptide or hit shows PDB IB, Classification, Experiment, Ligand, and Citations [63-69], see Table **2**. In each pdb can be downloaded and observed: Sequence FASTA, PDB File, File mmCIF, PDBML / XML File, Factor Structure, Biological Assembly (see Fig. **6**).

3. THEORETICAL STUDIES FOR SECRETASE INHIBITORS

3.1. Models of Novel Pyridinium-Based Potent β-Secretase Inhibitory Leads

In one article by Afaf Al-Nadaf *et al.* [70] explore the pharmacophoric space of 129 known BACE inhibitors have potential as anti-Alzheimer's disease treatments. The QSAR analysis employed to select optimal combination of pharmacophoric models and 2D physicochemical descriptors capable of explaining bioactivity variation (r^2 = 0.88, F = 60.48, r^2 LOO = 0.85, r^2 PRESS against 25 external test inhibitors = 0.71). They were obliged to use ligand efficiency as the response variable because the logarithmic transformation of bioactivities failed to access self-consistent QSAR models. The authors constructed three pharmacophoric models emerged in the successful QSAR equation suggesting at least three binding modes accessible to ligands within BACE binding pocket. The QSAR equation and pharmacophoric models were validated through ROC curves (see Table **3**), and were employed to guide synthesis of novel pyridinium-based BACE inhibitors.

3.2. CoMFA & CoMSIA of Hydroxyethylamine Derivatives as BACE-1 Inhibitors

Ashish Pandey *et al.* [71] were developed three-dimensional quantitative structure–activity relationship (3D-QSAR) models based on comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA), on a series of 43 hydroxyethylamine derivatives, acting as potent inhibitors of *β*-site amyloid precursor protein (APP) cleavage enzyme (BACE-1). They used a crystal structure of the BACE-1 enzyme (PDB ID: 2HM1) with one of the most active compound presented in this pa-

Table 1. ChEMBL Target Search Results for Secretase Enzymes

Target ID	Description	UniProt Ascension	Organism	Compounds	Endpoints
10231	γ-secretase subunit APH-1B	Q8WW43	Human	426	536
10322	γ -secretase subunit APH-1A	Q96BI3	Human	400	511
10378	Cathepsin B	P07858	Human	1431	1699
10656	B amyloid A4 protein	P05067	Human	361	561
10674	B secretase 2	Q9Y5Z0	Human	415	456
12252	B-secretase 1	P56817	Human	2157	3360
12711	γ -secretase subunit PEN-2	Q9NZ42	Human	495	621
100575	γ -secretase subunit PEN-2	Q9CQR7	Mouse	3	3
100580	γ -secretase subunit APH-1B	Q8C7N7	Mouse	3	3
100581	γ -secretase subunit APH-1A	Q8BVF7	Mouse	3	3
100609	B-secretase 1	P56818	Mouse	3	4

Fig. (5). Protein Data Bank server.

per was available, and we assumed it to be the bioactive conformation of the studied series, for 3D-QSAR analysis. Statistically significant 3D-QSAR model was established on a training set of 34 compounds, which were validated by a test set of 9 compounds. For the best CoMFA model, the statistics are, $r^2 = 0.998$, $r^2_{cv} = 0.810$, $n = 34$ for the training set and $r^2_{pred} = 0.934$, $n = 9$ for the test set. For the best CoMSIA model (combined steric, electrostatic, hydrophobic, and hydrogen bond donor fields), the statistics are $r^2 = 0.978$, $r^2_{cv} = 0.754$, $n = 34$ for the training set and $r^2_{pred} = 0.750$, $n = 9$ for the test set, see Table **4**. The resulting contour maps, produced by the best CoMFA and CoMSIA models, were used to identify the structural features relevant to the biological activity in series of analogs. The data generated from the present study will further help to design novel, potent, and selective BACE-1 inhibitors.

3.3. Virtual Screening and Protonation States at Asp32 and Asp228

György M. Keseru *et al.* [72] performed a comparative virtual screen for *β*-secretase (BACE1) inhibitors using different docking methods (FlexX and FlexX-Pharm), scoring functions (Dock, Gold, Chem, PMF, FlexX), protonation states (default and calculated), and protein conformations (apo and ligand bound). Apo and ligand bound conforma-

Fig. (6). PDB information about experimentally-determined structure for β-secretase.

tions of BACE1 were both found to be suitable for virtual screening. Assigning calculated protonation states to catalytic Asp32 and Asp228 residues resulted in significant improvement of enrichment factors as calculated at 1% of the ranked database. The authors used 1FKN to obtained no enrichment by FlexX/D-Score that was improved to ligand when considering calculated protonation states. They also show that combining calculated protonation states with pharmacophore constraints using FlexX-Pharm/D-Score improved enrichment further to ligand. Enrichments reported in this study suggest our screening protocol will be effective in the virtual screening of large compound libraries for BACE1 inhibitors.

3.4. Docking Scoring Function Based on 2D-Descriptors

In this paper Csaba Hetényi [73] showed a key step in the molecular engineering of such potent lead compounds is the prediction of the energetics of their binding to the macromolecular targets. Although sophisticated experimental and in silico methods are available to help this issue, the structure-based calculation of the binding free energies of large, flexible ligands to proteins is problematic. In this study, a fast and accurate calculation strategy is presented, following modification of the scoring function of the popular docking program package AutoDock and the involvement of ligand-based two-dimensional descriptors. Quantitative structure-activity relationships with good predictive power were developed. The best results of this paper were shown in Table **5.** Thorough cross-validation tests and verifications were performed on the basis of experimental binding data of biologically important systems. The capabilities and limitations of the ligand based descriptors were analyzed. According to the authors the application of these results in the early phase of lead design will contribute to precise predictions, correct selections, and consequently a higher success rate of rational drug discovery.

3.5. Induced-Fit Docking of Peptidic and Pseudo-Peptidic BACE-1 Inhibitors

Inhibition of *β*-secretase (BACE 1) has recently been investigated as a promising therapeutic approach inthe treatment of Alzheimer's disease, and a growing number of BACE 1 inhibitors and crystal structures of BACE 1/inhibitors complexes have been reported. Nicolas Moitessier *et al.* [74] report herein a predictive computational method and its application to potential BACE 1 inhibitors. Using a training set of 50 known highly flexible inhibitors, they developed a docking method that accounts for the flexibility of both the protein and the inhibitors. Protein flexibility is accounted for using a specifically designed genetic algorithm. In this paper developed a scoring function consisting of force field evaluation of the inhibitor/protein interactions and two additional terms for hydrogen bonding and entropy change upon binding. Discarding three outliers from the training set, the protocol was found to perform well with an rmsd of 1.19 kcal/mol an r^2 value of 0.789. Evaluation of the predictive power was carried out by virtual screening of 80 synthetic compounds. The significant enrichment at the top of the ranking list in active compounds demonstrated the ability of the docking and scoring protocol to rank the compounds relative to their activities.

3.6. Molecular Docking Studies of Phlorotannins BACE1 Inhibitory Activity

In your consecutive research on an anti-AD remedy derived from maritime plants, the BACE1 inhibitory activities of

Table 2. PDB Secretase Searcher Results for Structures Determined by X-Ray Method

PDB IB	Classification	Resolution (Å)	Ligand	Reference
3L81	Transport Protein	1.60	Glycerol	[63]
3I08	Signaling Protein	3.20	Ca^{2+}, Cl^-	[64]
3CBJ	Hidrolase inhibitor	1.80	Phosphate ion	-
2QP8	Hidrolase	1.50	D(-)-Tartaric acid	[65]
2V00	Hidrolase	1.55	Glycerol Acetate ion	[66]
2FJZ	Metal binding protein	1.61	-	[67]
1UJK	Protein transport hidrolase	1.90	Iodide ion	[70]
1BJB	Glycoprotein	1.61	-	[69]

Table 3. ROC_a Performances of QSAR-Selected Pharmacophores as 3D Search Queries

Pharmacophore	ROC_a/AUC_b	ACC_c	SPC_d	TPR_e	FNR_f
Hypo10/10	0.982	0.961	0.988	0.28	0.011345
Hypo6/18	0.981	0.961	0.975	0.6	0.024311
Hypo1/21	0.738	0.961	0.9611	0.96	0.038898

[a] ROC: receiver operating characteristic, [b] AUC: area under the curve, [c] ACC: overall accuracy,
[d] SPC: overall specificity, [e] TPR: overall true positive rate, [f] FNR: overall false negative rate.

Table 4. PLS Summary of CoMFA and CoMSIA Results

Statistical Parameters	CoMFA	CoMSIA
	(S E)	(S EHD)
Number of molecules in training set	34	34
Number of molecules in test set	9	9
r^2 cv	0.810	0.754
NOC	7	4
SEE	0.063	0.204
r^2	0.998	0.978
F-test	2009.08	324.673
r^2_{bs}	0.999	0.989
SD_{bs}	0.001	0.006
r^2_{pred}	0.934	0.750
Percentage of field contributions		
S	47.4	24.8
E	52.6	34.0
H	–	26.3
D	–	14.9

Abbreviations: S (steric field), E (electrostatic field), H (hydrophobic field), D (hydrogen bond donor field) r^2_{cv} =Cross-validated correlation coefficient by PLS LOO method, NOC=Optimum number of components as determined by PLS LOO cross-validation study, SEE=Standard error of estimate, r^2 =Conventional correlation coefficient, r^2_{bs} =Correlation coefficient after 100 runs of boot strapping, SD_{bs} =Standard deviation from 100 runs of bootstrapping, r^2_{pred} =Predictive correlation coefficient.

Table 5. The Results Produced by the Best CoMFA and CoMSIA Models

QSAR	Descriptor (Di)							
		coefficient (Ri)	error of coeff.	t-value	R^2	R^2_{cv}	s^2	F-value
A	1	¢GTH	3.1216 × 10-1	2.4686 × 10-2	0.799	0.774	1.05	93.36
	2	RPCGEN	3.2582 × 101	6.9963				
		constant	-4.1980	6.9930 × 10-1				
B	1	¢GTH	2.7077 × 10-1	2.2926 × 10-2	0.859	0.838	0.76	93.17
	2	RPCGEN	5.7129 × 101	8.1307				
	3	J	-6.2410 × 10-1	1.4148 × 10-1				
		constant	-4.6864	6.0281 × 10-1				

Standard deviations (s^2), squares of the correlation coefficients (R^2), and leave-one-out cross-validated correlation coefficients (R^2_{cv}) of the regressions are tabulated.

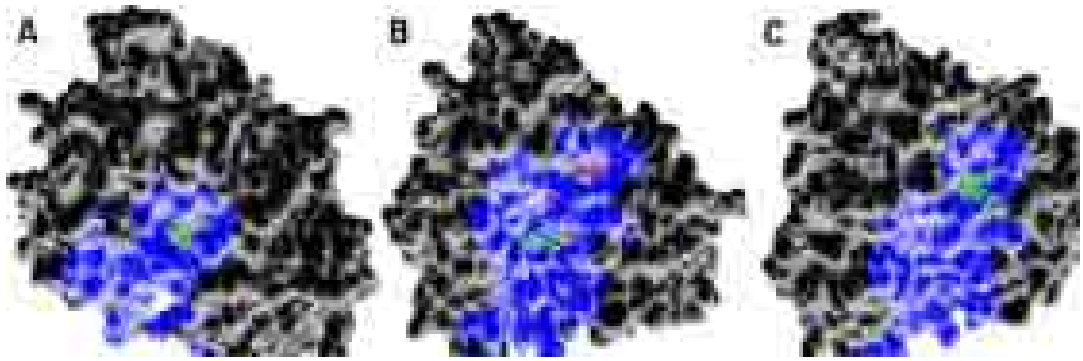


Fig. (7). Molecular docking models of BACE1-phlorotannins.

Eisenia bicyclis and its isolated phlorotannins were evaluated by Hyun Ah Jung *et al.* [75]. The E. bicyclis extract and its fractions exhibited predominant BACE1 inhibition. With the exception of one molecule of phloroglucinol, according to the authors all test phlorotannins isolated and showed significant and non-competitive inhibition against BACE1:dioxinodehydroeckol (2, IC_{50} = 5.35 lM; Ki = 8.0); eckol (3, IC_{50} = 12.20 lM; Ki = 13.9); phlorofurofucoeckol-A (4, IC_{50} = 2.13 lM; Ki = 1.3); dieckol (5, IC_{50} = 2.21 lM; Ki = 1.5); triphloroethol A (6, IC50 = 11.68 lM; Ki = 12.1); 7-phloroethol (7, IC_{50} = 8.59 lM; Ki = 7.2). In addition, plausible protein–ligand interactions of three molecules (eckol, dieckoland phlorofurofucoeckol) were similar and may occur primarily through the TYR132 and THR133 of BACE1 via molecular docking simulations (AUTODOCK 4.0 and FRED 2.0 programs) (see Fig. **7**). As a result, the E. bicyclis extract and three phlorotannins contained therein would clearly have beneficial use in the development of therapeutic and preventive agents for AD and suggest potential guidelines for the design of BACE selective inhibitors.

Table 6. CoMFA and COMSIA Results

	CoMFA	CoMSIA
PLS statistics		
q^2	0.582	0.622
r^2	0.986	0.982
S	0.091	0.105
F	307.229	191.762
Optimal	5	6
Field distribution (%)		
Steric	48.4	19.2
Electrostatic	51.6	47.5
Hydrophobic		33.3
Testing set		
r^2	0.756	0.853
S	0.237	0.225

3.7. Molecular Docking and 3D-QSAR Studies of Peptidomimetics with β-Secretase

β-Secretase is an important protease in the pathogenesis of Alzheimer´s disease. Some statine-based peptidomimetics show inhibitory activities to the β -secretase. Zhili Zuo *et al.*[76] were explore and performed the inhibitory mechanism, molecular docking and three-dimensional quantitative structure–activity relationship (3D-QSAR) studies on these analogues. The Lamarckian Genetic Algorithm method (LGA) was applied by the authors to locate the binding orientations and conformations of the peptidomimetics with the β-secretase. A good correlation between the calculated

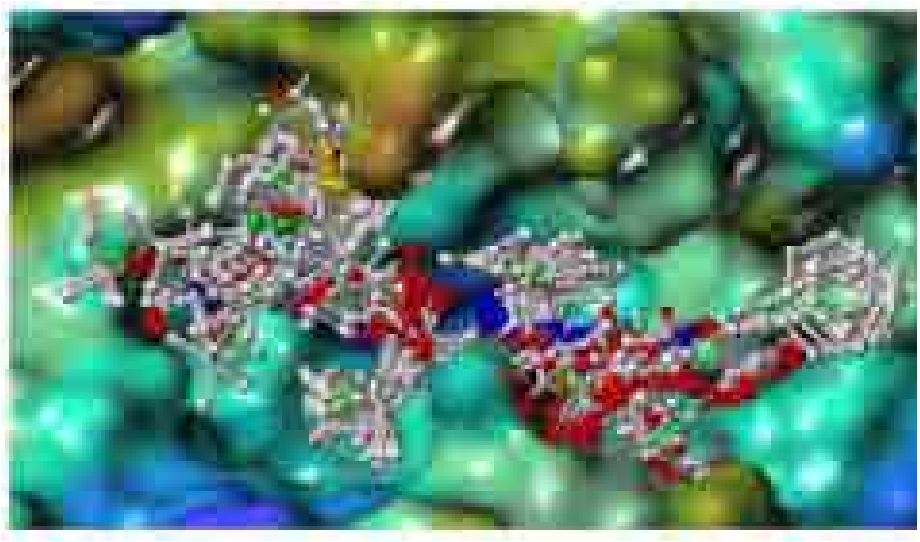


Fig. (8). The binding conformations of the statine-based compounds displayed inside the active site of the β-secretase.

binding free energies and the experimental inhibitory activities suggests that the identified binding conformations of these potential inhibitors are reliable. Based on the binding conformations, highly predictive 3D-QSAR models were developed with q^2 values of 0.582 and 0.622 for CoMFA and CoMSIA, respectively, all statistical results are in Table **6**. The predictive abilities of these models were validated by some compounds that were not included in the training set. Furthermore, the 3D-QSAR models were mapped back to the binding site of the β -secretase, to get a better understanding of vital interactions between the statine-based peptidomimetics and the protease. Based on the binding conformations from molecular docking, highly predictive CoMFA and CoMSIA models match well the 3D topology of the binding site of the b-secretase (see Fig. **8**). Therefore, the final 3D-QSAR models and the information of the inhibitor–enzyme interaction would be useful in developing new drug leads against Alzheimer´s disease.

4. NEW METHOD FOR THE STUDY OF NEW β-SECRETASE INHIBITORS

4.1. Preface to New ANN-QSAR Study β-Secretase Inhibitors

Alzheimer´s disease (AD) [77] is a serious and degenerative disorder that causes a the gradual loss of neurons, and in spite of the efforts realized by the big pharmaceutical companies of the world, the origen of this pathology is still not very clear. We can see in this paper that the development of theoretical and QSAR models to study β-secretase inhibitors are usually not many achieved so far, and most of these works present docking studies. Watching this situation we need to develop QSAR models with β-secretase inhibitors. In this sense, quantitative structure-activity relationships (QSAR) could play an important role in studying these β-secretase inhibitors; QSARs can be used as predictive tools for the development of molecules [78, 79]. Computer-aided drug design techniques based on Quantitative Structure-Activity Relationships (QSAR) could play an important role in drug discovery programs. The QSAR approach involves the development of models that relate the structure of drugs with their biological activity against different targets [80, 81]. In principle, there are currently more than 1600 molecular descriptors that may be generalized and used to solve the problem outlined above [82]. Numerous different molecular descriptors have been reported to encode chemical structures in QSAR studies. Furthermore, there are multiple chemometric approaches that can, in principle, be selected for this step. Multiple linear regression (MLR), linear discriminant analysis (LDA) [83], partial least squares (PLS) and different kinds of artificial neural networks (ANN) can be used to relate molecular structure (represented by molecular descriptors) with biological properties. The ANNs are particularly useful in QSAR studies in which the linear models fit poorly due to high data complexity [17, 18], an example was the work of Prado-Prado *et. al*. In which four types of non Linear Artificial neural networks (ANN) were developed for β-secretase inhibitors, ANNs was constructed from more than 15 000 cases with more than 3 000 different molecules inhibitors of β-secretase obtained from ChEMBL database http://
www.ebi.ac.uk/ChEMBLdb/index.php; in total we used more than 10 000 different molecules to develop the QSAR models. We used spectral moments molecular descriptors calculated with Modeslab software [61].

4.2. Method

4.2.1. Data Set

The data set used in this article was obtained from ChEMBL database. It has more than 20000 cases and more than 3 000 different compounds inhibitors of β-secretase. In total we used more than 10 000 different molecules to develop the QSAR models obtained in ChEMBL. This is a database of bioactive drug-like small molecules, it contains 2D structures, calculated properties (e.g. logP, Molecular Weight, Lipinski Parameters, etc.) and abstracted bioactivities (e.g. binding constants, pharmacology and ADMET data). ChEMBL normalises the bioactivities into a uniform set of end-points and units where possible, and also tags the links between a molecular target and a published assay with a set of varying confidence levels. The data is abstracted and curated from the primary scientific literature, and covers a significant fraction of the SAR and discovery of modern

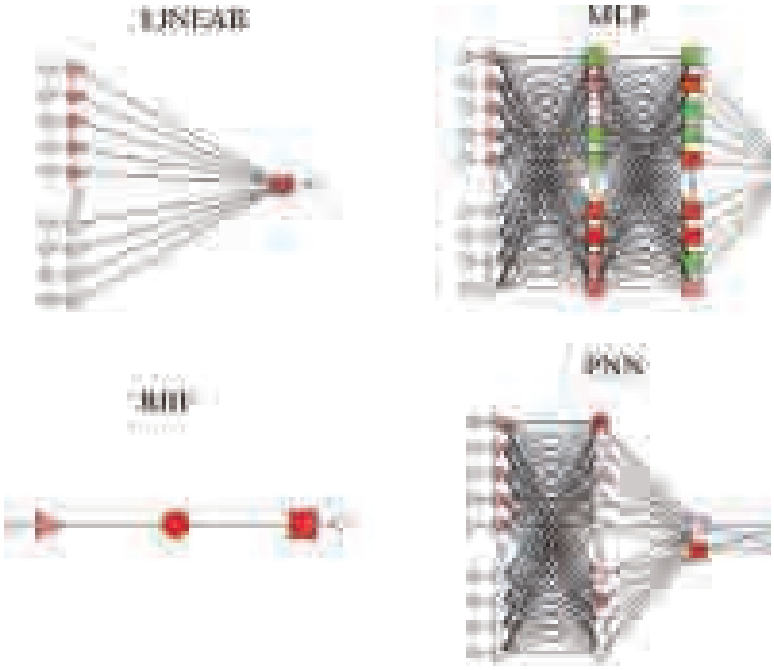


Fig. (9). Topology of some ANN models trained in this work.

Table 7. Comparison of Different ANNs Classification Models

Model		Train		Stat.		Validation	
profile	active	Non-active	%	Par.	active	Non-active	%
RBF	1255	849	59.65	Sn	628	425	59.64
1:1-326-1:1	2385	5915	71.27	Sp	1186	2964	71.42
			68.92	Ac			69.04
PNN	34	2070	1.62	Sn	17	1036	1.61
15:15-10404-2-2:1	0	8300	100	Sp	0	4150	100
			80.10	Ac			80.09
MLP	1812	292	86.12	Sn	909	144	86.32
15:15-11-1:1	1118	7182	86.53	Sp	556	3594	86.60
			86.45	Ac			86.55
Linear	1926	178	91.54	Sn	973	80	92.40
15:15-1:1	745	7555	91.02	Sp	308	3842	92.58
			91.13	Ac			92.54

drugs. The codes and activity for all compounds as well as the references used to collect them are depicted in **SM1** of the supplementary material file.

4.2.2. ANN Models

The ANN models are non-linear models useful to predict the biological activity of a large datasets of molecules. This technique is an alternative to linear methods such as LDA. (Fig. **9**) depicts the networks maps for some of the ANN models. In general, at least one ANN of every types tested was statically significant. However, one must note that the

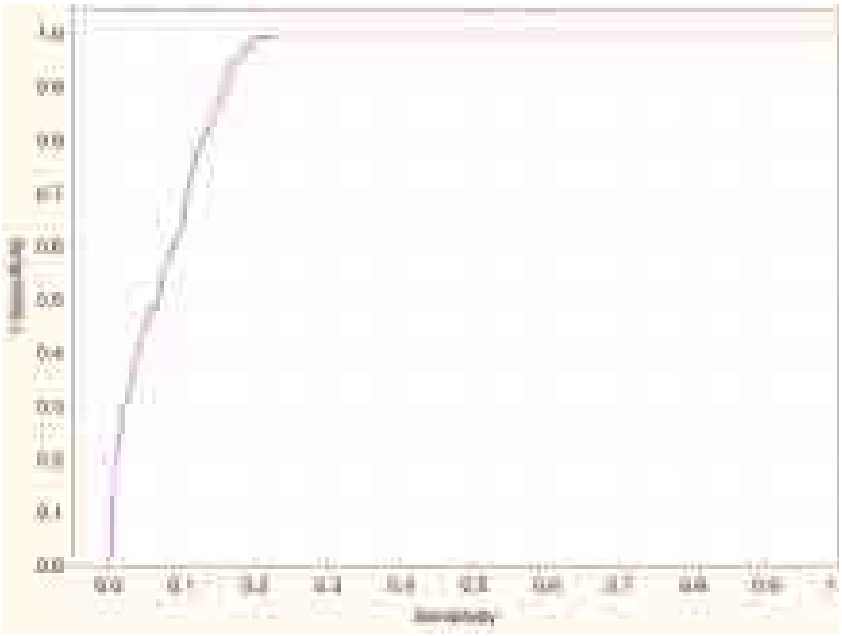


Fig. (10). ROC Curve for classifier.

profiles of each network indicate that these are highly nonlinear and complicated models.

There are several different kinds of ANN and these include multilayer perceptron (MLP), radial basis functions (RBF) and PNNs; the latter ANN is a variant of RBF systems. In particular, PNN is a type of neural network that uses a kernel-based approximation to form an estimate of the probability density functions of classes in a classification problem [84].

4.3. Results and Discussion

The network found was LNN and it showed training performance higher than 91%. We compare different types of networks to obtain a better model; Table **7** shows the classification matrix of the different networks. Linear 15:15-1:1 was taken as the main network because it presented a wider range of variables, 15 inputs in the first layer and 15 neurons in second layer, and two sets of cases (Training and Validation). Another tested networks found were MLP 15:15-11-1:1, RBF 1:1-326-1:1 presented low accuracy and PNN 15:15-10404-2-2:1 had a very low percentage of non-active leading to possible errors in the model although its accuracy very good, see Table **7**. In Fig. (**10**), we depict the ROC-curve [85] for LNN tested. Notably, almost model presented and an area under curve higher than 0.5 (the value for a random classifier). The vitality of this type of procedures developing ANN-QSAR models has been demonstrated befote [86]; see, for instance, the work of Fernandez and Caballero [87]. The same is true about the ANNs tested, where is illustrated ROC-curve of ANN LNN with an area higher than 0.93. To show how important is this result, we compared the present model with other model used to address the same problem. We processed our data with Artificial Neural Networks (ANNs) looking for a better model. In general, the ANN LNN tested was statically significant.

5. CONCLUSIONS

Theoretical studies such as QSAR models have become a very useful tool in this context to substantially reduce time and resources consuming experiments. The functions of β-secretase and its implication in Alzheimer's disease have triggered an active search for potent and selective β-secretase inhibitors. In this paper we can see that the development of theoretical and QSAR models to study β-secretase inhibitors are usually not many achieved so far, and most of these works present docking studies. Watching this situation we need to develop QSAR models with β-secretase inhibitors. In this sense, QSAR could play an important role in studying these β-secretase inhibitors. QSARs can be used as predictive tools for the development of molecules. In this work we developed a new ANN LNN model using the ModesLab descriptors, based on a large database using about 10,000 different drugs obtained from the ChEMBL server.

ACKNOWLEDGEMENTS

F. Prado-Prado thanks sponsorships for research position at the University of Santiago de Compostela from *Angeles Alvariño*, Xunta de Galicia. All authors acknowledge the Project 07CSA008203PR.

REFERENCES

[1] Querfurth HW, LaFerla FM. Alzheimer's disease. N Engl J Med 2010; 362: 329-44.

[2] Goate A, Chartier-Harlin MC, Mullan M, *et al.* Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature 1991; 349: 704-6.

[3] Levy-Lahad E, Wasco W, Poorkaj P, *et al.* Candidate gene for the chromosome 1 familial Alzheimer's disease locus. Science 1995; 269: 973-7.

[4] Sherrington R, Rogaev EI, Liang Y, *et al.* Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. Nature 1995; 375: 754-60.

[5] Citron M, Westaway D, Xia W, *et al.* Mutant presenilins of Alzheimer's disease increase production of 42-residue amyloid beta-protein in both transfected cells and transgenic mice. Nat Med 1997; 3: 67-72.

[6] Holcomb L, Gordon MN, McGowan E, *et al.* Accelerated Alzheimer-type phenotype in transgenic mice carrying both mutant

amyloid precursor protein and presenilin 1 transgenes. Nat Med 1998; 4: 97-100.
[7] Cole SL, Vassar R. The Alzheimer's disease beta-secretase enzyme, BACE1. Mol Neurodegener 2007; 2: 22.
[8] Cole SL, Vassar R. The basic biology of BACE1: a key therapeutic target for alzheimer's disease. Curr Genomics 2007; 8: 509-30.
[9] Haass C, Schlossmacher MG, Hung AY, *et al.* Amyloid beta-peptide is produced by cultured cells during normal metabolism. Nature 1992; 359: 322-5.
[10] Seubert P, Oltersdorf T, Lee MG, *et al.* Secretion of beta-amyloid precursor protein cleaved at the amino terminus of the beta-amyloid peptide. Nature 1993; 361: 260-3.
[11] Fan Y, Unwalla R, Denny RA, *et al.* Insights for predicting blood-brain barrier penetration of CNS targeted molecules using QSPR approaches. J Chem Inf Model 2010; 50: 1123-33.
[12] Chou KC. Structural bioinformatics and its impact to biomedical science. Curr Med Chem 2004; 11: 2105-34.
[13] Chou KC, Wei DQ, Du QS, Sirois S, Zhong WZ. Progress in computational approach to drug development against SARS. Curr Med Chem 2006; 13: 3263-70.
[14] Todeschini R, Consonni V. Handbook of Molecular Descriptors. Wiley-VCH 2002.
[15] Estrada E, Uriarte E. Recent advances on the role of topological indices in drug discovery research. Curr Med Chem 2001; 8: 1573-88.
[16] González-Díaz H, González-Díaz Y, Santana L, Ubeira FM, Uriarte E. Proteomics, networks and connectivity indices. Proteomics 2008; 8: 750-78.
[17] Nunez MB, Maguna FP, Okulik NB, Castro EA. QSAR modeling of the MAO inhibitory activity of xanthones derivatives. Bioorg Med Chem Lett 2004; 14: 5611-7.
[18] Terada M, Inaba M, Yano Y, *et al.* Growth-inhibitory effect of a high glucose concentration on osteoblast-like cells. Bone 1998; 22: 17-23.
[19] Santana L, Uriarte E, González-Díaz H, Zagotto G, Soto-Otero R, Mendez-Alvarez E. A QSAR model for in silico screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins. J Med Chem 2006; 49: 1149-56.
[20] Marrero-Ponce Y. Linear indices of the "molecular pseudograph's atom adjacency matrix": definition, significance-interpretation, and application to QSAR analysis of flavone derivatives as HIV-1 integrase inhibitors. J Chem Inf Comput Sci 2004; 44: 2010-26.
[21] Vilar S, Santana L, Uriarte E. Probabilistic neural network model for the in silico evaluation of anti-HIV activity and mechanism of action. J Med Chem 2006; 49: 1118-24.
[22] Marrero-Ponce Y, Khan MT, Casanola Martin GM, *et al.* Prediction of Tyrosinase Inhibition Activity Using Atom-Based Bilinear Indices. ChemMedChem 2007; 2: 449-78.
[23] Casanola-Martin GM, Marrero-Ponce Y, Khan MT, *et al.* TOMO-COMD-CARDD descriptors-based virtual screening of tyrosinase inhibitors: evaluation of different classification model combinations using bond-based linear indices. Bioorg Med Chem 2007; 15: 1483-503.
[24] Casanola-Martin GM, Marrero-Ponce Y, Khan MT, *et al.* Dragon method for finding novel tyrosinase inhibitors: Biosilico identification and experimental *in vitro* assays. Eur J Med Chem 2007; 42: 1370-81.
[25] Caballero J, Fernandez M. Artificial neural networks from MATLAB in medicinal chemistry. Bayesian-regularized genetic neural networks (BRGNN): application to the prediction of the antagonistic activity against human platelet thrombin receptor (PAR-1). Curr Top Med Chem 2008; 8: 1580-605.
[26] Duardo-Sanchez A, Patlewicz G, Lopez-Diaz A. Current topics on software use in medicinal chemistry: intellectual property, taxes, and regulatory issues. Curr Top Med Chem 2008; 8: 1666-75.
[27] Gonzalez MP, Teran C, Saiz-Urra L, Teijeira M. Variable selection methods in QSAR: an overview. Curr Top Med Chem 2008; 8: 1606-27.
[28] Gonzalez-Diaz H. Quantitative studies on Structure-Activity and Structure-Property Relationships (QSAR/QSPR). Curr Top Med Chem 2008; 8: 1554.
[29] Gonzalez-Diaz H, Prado-Prado F, Ubeira FM. Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. Curr Top Med Chem 2008; 8: 1676-90.
[30] Helguera AM, Combes RD, Gonzalez MP, Cordeiro MN. Applications of 2D descriptors in drug design: a DRAGON tale. Curr Top Med Chem 2008; 8: 1628-55.
[31] Ivanciuc O. Weka machine learning for predicting the phospholipidosis inducing potential. Curr Top Med Chem 2008; 8: 1691-709.
[32] Vilar S, Cozza G, Moro S. Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. Curr Top Med Chem 2008; 8: 1555-72.
[33] Wang JF, Wei DQ, Chou KC. Drug candidates from traditional chinese medicines. Curr Top Med Chem 2008; 8: 1656-65.
[34] Wang JF, Wei DQ, Chou KC. Pharmacogenomics and personalized use of drugs. Curr Top Med Chem 2008; 8: 1573-9.
[35] Torrens F, Castellano G. Topological charge-transfer indices: from small molecules to proteins. Curr Proteomics 2009: 204-13.
[36] Concu R, Dea-Ayuela MA, Perez-Montoto LG, *et al.* 3D entropy and moments prediction of enzyme classes and experimental-theoretic study of peptide fingerprints in Leishmania parasites. Biochimica et Biophysica Acta 2009; 1794: 1784-94.
[37] Ivanciuc O. Machine learning Quantitative Structure-Activity Relationships (QSAR) for peptides binding to human amphiphysin-1 SH3 domain. Curr Proteomics 2009; 4: 289-302.
[38] Vilar S, Gonzalez-Diaz H, Santana L, Uriarte E. A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer. J Theor Biol 2009; 261: 449-58.
[39] Giuliani A, Di Paola L, Setola R. Proteins as networks: a mesoscopic approach using haemoglobin molecule as case study. Curr Proteomics 2009; 6: 235-45.
[40] Chou KC. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr Proteomics 2009; 6: 262-74.
[41] Chen J, Shen B. Computational analysis of amino acid mutation: a proteome wide perspective. Curr Proteomics 2009; 6: 228-34.
[42] Zhong WZ, Zhan J, Kang P, Yamazaki S. Gender specific drug metabolism of PF-02341066 in rats--role of sulfoconjugation. Curr Drug Metab 2010; 11: 296-306.
[43] Wang JF, Chou KC. Molecular modeling of cytochrome P450 and drug metabolism. Curr Drug Metab 2010; 11: 342-6.
[44] Mrabet Y, Semmar N. Mathematical methods to analysis of topology, functional variability and evolution of metabolic systems based on different decomposition concepts. Curr Drug Metab 2010; 11: 315-41.
[45] Martinez-Romero M, Vazquez-Naya JM, Rabunal JR, *et al.* Artificial intelligence techniques for colorectal cancer drug metabolism: ontology and complex network. Curr Drug Metab 2010; 11: 347-68.
[46] Khan MT. Predictions of the ADMET properties of candidate drug molecules utilizing different QSAR/QSPR modelling approaches. Curr Drug Metab 2010; 11: 285-95.
[47] Gonzalez-Diaz H, Duardo-Sanchez A, Ubeira FM, *et al.* Review of MARCH-INSIDE & complex networks prediction of drugs: ADMET, anti-parasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers. Curr Drug Metab 2010; 11: 379-406.
[48] Gonzalez-Diaz H. Network topological indices, drug metabolism, and distribution. Curr Drug Metab 2010; 11: 283-4.
[49] Garcia I, Diop YF, Gomez G. QSAR & complex network study of the HMGR inhibitors structural diversity. Curr Drug Metab 2010; 11: 307-14.
[50] Chou KC. Graphic rule for drug metabolism systems. Curr Drug Metab 2010; 11: 369-78.
[51] Concu R, Podda G, Ubeira FM, Gonzalez-Diaz H. Review of QSAR models for enzyme classes of drug targets: Theoretical background and applications in parasites, hosts, and other organisms. Curr Pharm Design 2010; 16: 2710-23.
[52] Estrada E, Molina E, Nodarse D, Uriarte E. Structural contributions of substrates to their binding to P-Glycoprotein. A TOPS-MODE approach. Curr Pharm Design 2010; 16: 2676-709.
[53] Garcia I, Fall Y, Gomez G. QSAR, docking, and CoMFA studies of GSK3 inhibitors. Curr Pharm Design 2010; 16: 2666-75.
[54] Gonzalez-Diaz H. QSAR and complex networks in pharmaceutical design, microbiology, parasitology, toxicology, cancer, and neurosciences. Curr Pharm Design 2010; 16: 2598-600.
[55] Gonzalez-Diaz H, Romaris F, Duardo-Sanchez A, *et al.* Predicting drugs and proteins in parasite infections with topological indices of complex networks: theoretical backgrounds, applications, and legal issues. Curr Pharm Design 2010; 16: 2737-64.

[56] Marrero-Ponce Y, Casanola-Martin GM, Khan MT, Torrens F, Rescigno A, Abad C. Ligand-based computer-aided discovery of tyrosinase inhibitors. Applications of the TOMOCOMD-CARDD method to the elucidation of new compounds. Curr Pharm Design 2010; 16: 2601-24.

[57] Munteanu CR, Fernandez-Blanco E, Seoane JA, *et al.* Drug discovery and design for complex diseases through QSAR computational methods. Curr Pharm Design 2010; 16: 2640-55.

[58] Roy K, Ghosh G. Exploring QSARs with Extended Topochemical Atom (ETA) indices for modeling chemical and drug toxicity. Curr Pharm Design 2010; 16: 2625-39.

[59] Speck-Planche A, Scotti MT, de Paulo-Emerenciano V. Current pharmaceutical design of antituberculosis drugs: future perspectives. Curr Pharm Design 2010; 16: 2656-65.

[60] Vazquez-Naya JM, Martinez-Romero M, Porto-Pazos AB, *et al.* Ontologies of drug discovery and design for neurology, cardiology and oncology. Curr Pharm Design 2010; 16: 2724-36.

[61] Estrada E. On the topological sub-structural molecular design (TOSS-MODE) in QSPR/QSAR and drug design research. SAR QSAR Environ Res 2000; 11: 55-73.

[62] Overington J. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. J Comput Aided Mol Des 2009; 23: 195-8.

[63] Burgos PV, Mardones GA, Rojas AL, *et al.* Sorting of the Alzheimer's disease amyloid precursor protein mediated by the AP-4 complex. Dev Cell 2010; 18: 425-36.

[64] Gordon WR, Vardar-Ulu D, L'Heureux S, *et al.* Effects of S1 cleavage on the structure, surface export, and signaling activity of human Notch1 and Notch2. PLoS One 2009; 4: e6613.

[65] Iserloh U, Wu Y, Cumming JN, *et al.* Potent pyrrolidine- and piperidine-based BACE-1 inhibitors. Bioorg Med Chem Lett 2008; 18: 414-7.

[66] Geschwindner S, Olsson LL, Albert JS, *et al.* Discovery of a novel warhead against beta-secretase through fragment-based lead generation. J Med Chem 2007; 50: 5903-11.

[67] Kong GK, Adams JJ, Harris HH, *et al.* Structural studies of the Alzheimer's amyloid precursor protein copper-binding domain reveal how it binds copper ions. J Mol Biol 2007; 367: 148-61.

[68] Shiba T, Kametaka S, Kawasaki M, *et al.* Insights into the phosphoregulation of beta-secretase sorting signal by the VHS domain of GGA1. Traffic 2004; 5: 437-48.

[69] Poulsen SA, Watson AA, Fairlie DP, Craik DJ. Solution structures in aqueous SDS micelles of two amyloid beta peptides of A beta(1-28) mutated at the alpha-secretase cleavage site (K16E, K16F). J Struct Biol 2000; 130: 142-52.

[70] Al-Nadaf A, Abu Sheikha G, Taha MO. Elaborate ligand-based pharmacophore exploration and QSAR analysis guide the synthesis of novel pyridinium-based potent beta-secretase inhibitory leads. Bioorg Med Chem 2010; 18: 3088-115.

[71] Pandey A, Mungalpara J, Mohan CG. Comparative molecular field analysis and comparative molecular similarity indices analysis of hydroxyethylamine derivatives as selective human BACE-1 inhibitor. Mol Divers 2010; 14: 39-49.

[72] Polgar T, Keseru GM. Virtual screening for beta-secretase (BACE1) inhibitors reveals the importance of protonation states at Asp32 and Asp228. J Med Chem 2005; 48: 3749-55.

[73] Hetenyi C, Paragi G, Maran U, Timar Z, Karelson M, Penke B. Combination of a modified scoring function with two-dimensional descriptors for calculation of binding affinities of bulky, flexible ligands to proteins. J Am Chem Soc 2006; 128: 1233-9.

[74] Moitessier N, Therrien E, Hanessian S. A method for induced-fit docking, scoring, and ranking of flexible ligands. Application to peptidic and pseudopeptidic beta-secretase (BACE 1) inhibitors. J Med Chem 2006; 49: 5885-94.

[75] Jung HA, Oh SH, Choi JS. Molecular docking studies of phlorotannins from Eisenia bicyclis with BACE1 inhibitory activity. Bioorg Med Chem Lett 2010; 20: 3211-5.

[76] Zuo Z, Luo X, Zhu W, *et al.* Molecular docking and 3D-QSAR studies on the binding mechanism of statine-based peptidomimetics with beta-secretase. Bioorg Med Chem 2005; 13: 2121-31.

[77] Salmon SA, Watts JL. Minimum inhibitory concentration determinations for various antimicrobial agents against 1570 bacterial isolates from turkey poults. Avian Dis 2000; 44: 85-98.

[78] Chou KC. Review: Structural bioinformatics and its impact to biomedical science. Curr Med Chem 2004; 11: 2105-34.

[79] Chou KC, Wei DQ, Du QS, Sirois S, Zhong WZ. Review: progress in computational approach to drug development against SARS. Curr Med Chem 2006; 13: 3263-70.

[80] Prado-Prado FJ, Gonzalez-Diaz H, de la Vega OM, Ubeira FM, Chou KC. Unified QSAR approach to antimicrobials. Part 3: first multi-tasking QSAR model for input-coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. Bioorg Med Chem 2008; 16: 5871-80.

[81] Prado-Prado FJ, de la Vega OM, Uriarte E, Ubeira FM, Chou KC, Gonzalez-Diaz H. Unified QSAR approach to antimicrobials. 4. Multi-target QSAR modeling and comparative multi-distance study of the giant components of antiviral drug-drug complex networks. Bioorg Med Chem 2009; 17: 569-75.

[82] Kubinyi H. Quantitative structure-activity relationships (QSAR) and molecular modelling in cancer research. J Cancer Res Clin Oncol 1990; 116: 529-37.

[83] Prado-Prado FJ, Borges F, Perez-Montoto LG, Gonzalez-Diaz H. Multi-target spectral moment: QSAR for antifungal drugs vs. different fungi species. Eur J Med Chem 2009; 44(10): 4051-6.

[84] Mosier PD, Jurs PC. QSAR/QSPR studies using probabilistic neural networks and generalized regression neural networks. J Chem Inf Comput Sci 2002; 42: 1460-70.

[85] Lombardi G, Gramegna G, Cavanna C, Michelone G. Fluconazole vs amphotericin B: "*in vitro*" comparative evaluation of the minimal inhibitory concentration (MIC) against yeasts isolated from AIDS patients. Microbiologica 1990; 13: 201-6.

[86] Prado-Prado FJ, Garcia-Mera X, Gonzalez-Diaz H. Multi-target spectral moment QSAR versus ANN for antiparasitic drugs against different parasite species. Bioorg Med Chem 2010; 18: 2225-31.

[87] Fernandez M, Caballero J, Tundidor-Camba A. Linear and nonlinear QSAR study of N-hydroxy-2-[(phenylsulfonyl)amino] acetamide derivatives as matrix metalloproteinase inhibitors. Bioorg Med Chem 2006; 14: 4137-50.

Received: October 10, 2010 Revised: October 20, 2010 Accepted: December 09, 2010

Review of Synthesis, Biological Assay and QSAR Studies of β-Secretase Inhibitors

Helena Niño[1], Isela García-Pintos[1], José E. Rodríguez-Borges[2], Manolo Escobar-Cubiella[1], Xerardo García-Mera[1] and Francisco Prado-Prado[*,1]

[1]*Department of Organic Chemistry, University of Santiago de Compostela, Santiago de Compostela, Spain*

[2]*CIQ-Departamento de Química, Faculdade de Ciencias, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007, Porto, Portugal*

Abstract: Alzheimer's disease (AD) is highly complex. While several pathologies characterize this disease, amyloid plaques, composed of the β-amyloid peptide, are hallmark neuropathological lesions in Alzheimer's disease brain. Indeed, a wealth of evidence suggests that β-amyloid is central to the pathophysiology of AD and is likely to play an early role in this intractable neurodegenerative disorder. The BACE-1 enzyme is essential for the generation of β-amyloid. BACE-1 knockout mice do not produce β-amyloid and are free from Alzheimer's associated pathologies, including neuronal loss and certain memory deficits. The fact that BACE-1 initiates the formation of β-amyloid, and the observation that BACE-1 levels are elevated in this disease provide direct and compelling reasons to develop therapies directed at BACE-1 inhibition, thus reducing β-amyloid and its associated toxicities. In this sense, quantitative structure-activity relationships (QSAR) could play an important role in studying these *β*-secretase inhibitors. QSAR models are necessary in order to guide the *β*-secretase synthesis. This work is aimed at reviewing different design and synthesis and computational studies for a very large and heterogeneous series of *β*-secretase inhibitors. First, we review design, synthesis, and Biological assay of *β*-secretase inhibitors. Next, we review 2D QSAR, 3D QSAR, CoMFA, CoMSIA and Docking with different compounds to find out the structural requirements. Next, we review QSAR studies using the method of Linear Discriminant Analysis (LDA) in order to understand the essential structural requirement for receptor binding for *β*-secretase inhibitors.

Keywords: QSAR, CoMSIA, COMFA, topological indices, *β*-secretase inhibitors, Alzheimer's disease (AD).

INTRODUCTION

AD is the most prevalent form of dementia, and current indications show that twenty-nine million people live with AD worldwide, a figure expected to rise exponentially over the coming decades. Obviously blocking disease progression or, in the best-case scenario, preventing AD altogether would be of benefit in both social and economic terms. While familial AD (FAD) is caused by autosomal dominant mutations in either amyloid precursor protein (APP) [1] or the presenilin (PS1, PS2) [2] genes, the underlying cause (s) of the remaining ~98% of so-called sporadic AD (SAD) cases remains elusive.

Pathologically, AD is characterized by the accumulation of amyloid beta peptide (Aβ), as fibrillar plaques and soluble oligomers in high-order association brain regions. The presence of intracellular neurofibrillary tangles, neuroinflammation, neuronal dysfunction and death further characterize this disease. Mounting evidence suggests that Aβ plays a critical early role in AD pathogenesis, and the basic tenant/tenet of the amyloid (or Aβ cascade) hypothesis is that Aβ aggregates trigger a complex pathological cascade which leads to neurodegeneration [3]. There is a strong genetic correlation between FAD and the 42 amino acid Aβ form (Aβ42; reviewed in [4-6]). Aβ is derived from APP and mutations in APP and PS increase Aβ42 production and cause FAD with nearly 100% penetrance. Down's syndrome (DS) patients, who have an extra copy of the APP gene on chromosome 21, and FAD families with a duplicated APP gene locus [7], exhibit total Aβ overproduction and all develop early-onset AD. In FAD, the Aβ42 increase is present years before AD symptoms arise, suggesting that Aβ42 is likely to initiate AD pathophysiology. The robust association of Aβ42 overproduction with FAD argues strongly in favor of a critical role for Aβ42 in the etiology of AD, including in SAD. Fibrillar and oligomeric forms of Aβ appear neurotoxic *in vitro* and *in vivo*. Importantly, in specific transgenic (Tg) mouse models of AD the lack of Aβ correlates with the absence of neuronal loss and improved cognitive function [8-10]. Such data provide direct evidence for the amyloid hypothesis *in vivo*, and also indicate that Aβ is directly responsible for neuronal death. Consequently, strategies to lower Aβ42 levels in the brain are anticipated to be of therapeutic benefit in AD.

Aβ peptide is generated following the sequential cleavage of APP by β- and γ-secretase in the amyloidogenic pathway [11, 12]. Aβ genesis may be precluded if APP is cleaved by α-secretase within the Aβ domain in the non-amyloidogenic pathway Fig. (**1**). Recently, the secretases have been identified and the β-secretase is known to be β-site APP cleaving enzyme 1 (BACE-1) [13-16], a novel aspartyl protease. BACE-1 cleavage of APP is a pre-requisite for Aβ formation. Aβ genesis is initiated by BACE-1 cleavage of APP at the Asp+1 residue of the Aβ sequence

*Address correspondence to this author at the Department of Organic Chemistry, University of Santiago de Compostela, Santiago de Compostela, Spain; Tel: 881814993; Fax: 981594912; E-mail: francisco.prado@usc.es

1573-4099/11 $58.00+.00

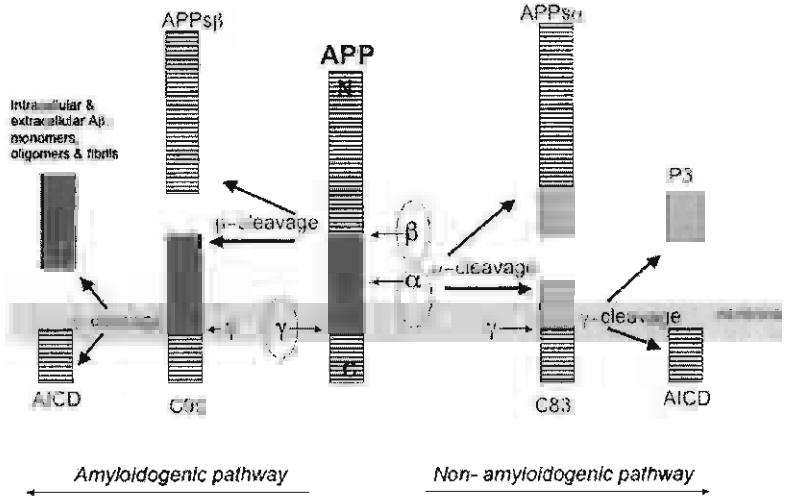


Fig. (1).

to form the N-terminus of the peptide. This scission liberates two cleavage fragments: a secreted APP ectodomain, APPsβ and a membrane-bound carboxyl terminal fragment (CTF). In many instances, an increase in non-amyloidogenic APP metabolism is coupled to a reciprocal decrease in the amyloidogenic processing pathway, and vice-versa, as the α- and β-secretase moieties compete for APP substrate [12, 15].

Given that BACE-1 is the initiating enzyme in Aβ generation, and putatively rate-limiting, it is considered a prime drug target for lowering cerebral Aβ levels in the treatment and/or prevention of AD. Prior to its identification, numerous studies were undertaken to define the characteristics of β-secretase activity. Although the majority of body tissues exhibit β-secretase activity [17], the highest activity levels were observed in neural tissue and neuronal cell lines [18]. Indeed, β-secretase appeared to predominate in neurons, with the level of β-secretase activity appearing lower in astrocytes [19]. Data showing that β-secretase efficiently cleaved only membrane-bound substrates [20] indicated that the enzyme was likely membrane-bound or closely associated with a membrane protein. Drugs that block this enzyme (BACE inhibitors) in theory would prevent the build-up of beta-amyloid and may help slow or stop the disease. However, current AD therapies are merely palliative and only temporarily slow cognitive decline, and treatments that address the underlying pathologic mechanisms of AD are completely lacking. In this sense, quantitative structure-activity relationships (QSAR) could play an important role in studying these *β*-secretase inhibitors. QSAR models are necessary in order to guide the *β*-secretase synthesis.

On the other hand, QSAR models can be used to explore the relationships between the structural spaces of compounds as inhibitors for specific enzymes, such as MAO inhibitors [21], HIV-1 integrase inhibitors [22], and/or protease inhibitors [23] or tyrosinase inhibitors [24-26]. In fact, almost all QSAR techniques are based on the use of molecular descriptors, which are numerical series that codify useful chemical information and enable correlations between statistical and biological properties [27, 28]. Recently, the field has moved from small molecules to proteins and other systems. For instance, González-Díaz *et al.* discussed the use of these methods, but referring only proteins [29]. Later, some groups published different papers in one special issue on QSAR, but they were also restricted to the field of proteins and proteomics [30-36]. In other recent issue, guest-edited by González-Díaz [37] a series of papers devoted to QSAR/QSPR techniques for low-molecular-weight drugs [37-46] was published. Most recently, Prado-Prado *et al.* [47] have published an mt-QSAR for anti-parasitic drugs. This year, another issue [48] has been published, focusing on QSAR/QSPR models and graph theory used to approach Drug ADMET processes and Metabolomics [49-56]. Last, one of the most recent issues published has been devoted to discuss the applications of QSAR in Pharmaceutical Design [57-66]. In the present work, we firstly review the state-of-the-art on the design, synthesis, and biological assay of *β*-secretase inhibitors. Next, we review previous works based on 2D-QSAR, 3D-QSAR, CoMFA, CoMSIA and Docking techniques, which studied different compounds to find out the structural requirements. Last, we carry out new QSAR studies using the Linear Discriminant Analysis (LDA) method and the software ModesLab in order to understand the essential structural requirement for receptor binding for *β*-secretase inhibitors. The reviewed, discussed, and/or reported topics in this paper are the following:

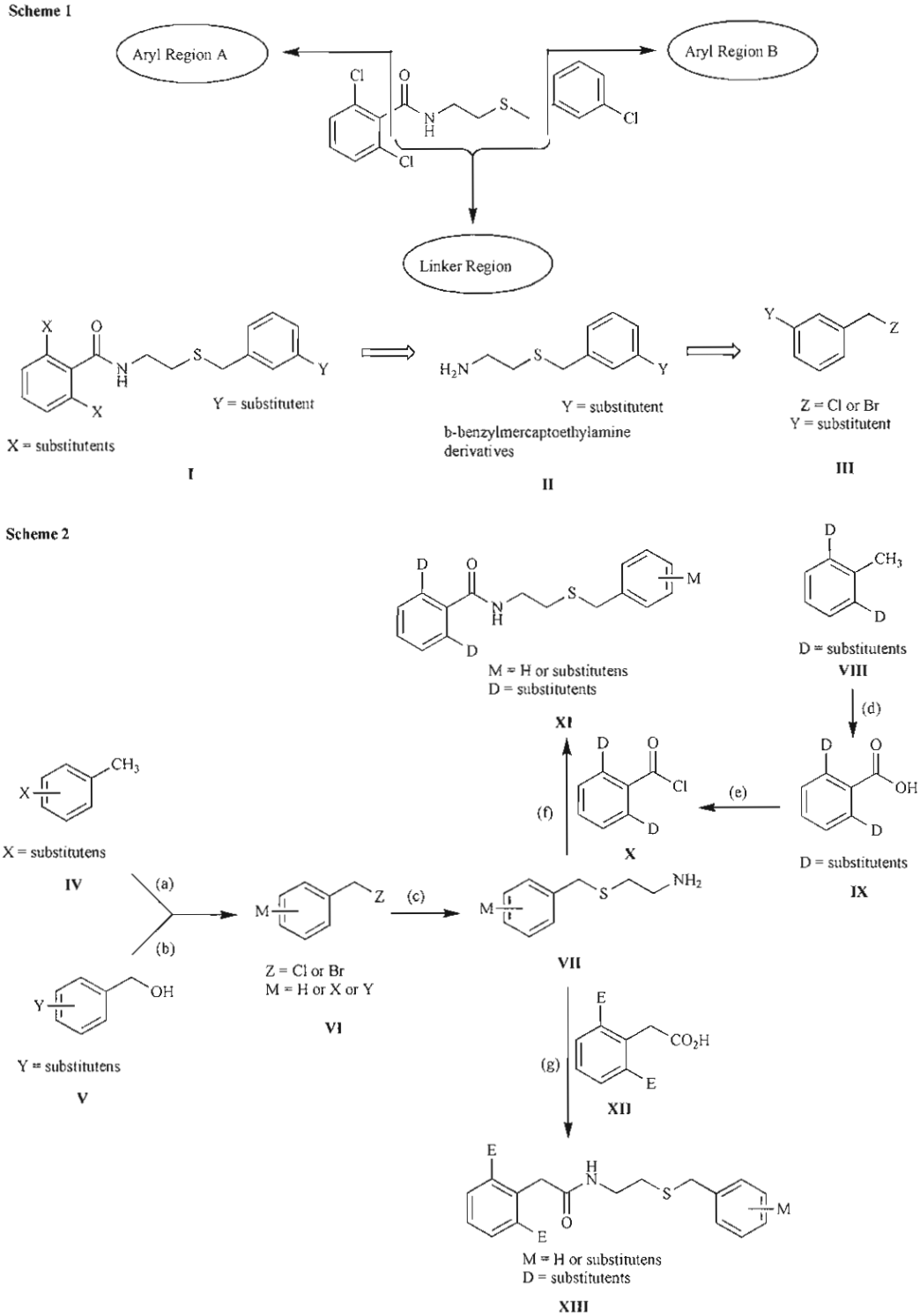
DISCUSSION

1. Synthesis and Biological Assay of β-Secretase Inhibitors

In this section, we have compiled some papers about the synthesis of β-secretase inhibitors. These compounds promoted the synthesis and evaluation of new inhibitors for β-secretase.

1.1. Synthesis and Assay of Novel Inhibitors of β-Amyloid Secretion

Enakshi Chakrabarti *et al.* [67] screened a drug library of 17200 compounds to select small molecules that inhibit the secretion of amyloid β peptide (Aβ), the major component of Alzheimer's disease senile plaques, from a human neuronal cell line. In this work, they validated twenty-nine hits that decreased Aβ secretion by > 40% at 10 μM, for a 0.17% hit rate. A lead hit was selected for further studies based on its activity and low cytotoxicity, and it was found to inhibit Aβ secretion through activation of the R-secretase pathway.



Reagents: (a) NBS, BPO, CCl_4, reflux; (b) $SOCl_2$, Et_3N, DCM, 50°C; (c) LiOH-H_2O-EtOH, $HSCH_2CH_2NH_2$ 3 HCl, 35°C, 40 min; (d) $KMnO_4$, t-BuOH-H_2O, celite, reflux, conc HCl; (e) $SOCl_2$, reflux; (f) Et_3N, DCM, 0°C to RT; (g) DCC, Et_3N, DCM, rt.

Fig. (2). Scheme **1**: Synthetic Analysis for Defining a Structure Activity Relationship for 1. Scheme **2**: General Synthetic Scheme for the Preparation of 1 Analogues[a].

Twenty-four commercially available and 53 synthesized analogues were analyzed for activity, as shown in Fig. (**2**). For biological stability, the selected analogues were evaluated by incubation with hepatoma cells and for trans-cellular permeability using Caco-2 cell mono-layers.

1.2. Design and Synthesis of Aminohydantoins as BACE-1 Inhibitors

Michael S. Malamas *et al.* [68] reported the synthesis and identification of small molecule aminohydantoins as potent and selective human β-secretase inhibitors (Fig. **3**). These analogues exhibit low nanomolar potency for BACE-1, show comparable activity in a cell-based (ELISA) assay, and demonstrate >100_ selectivity for the other structurally-related aspartyl proteases BACE2, cathepsin D, renin, and pepsin. On the basis of the cocrystal structure of the HTS-hit 2 in the BACE-1 active site and using a structure-based drug design approach, in this paper the authors explored methodically the comparatively large binding pocket of the BACE-1 enzyme and identified key interactions between the ligand and the protein that contributed to the affinity. One of the most potent compounds, (S)-55, displayed an IC50 value for BACE-1 of 10 nM and exhibited comparable cellular activity (EC50=20 nM) in the ELISA assay. Acute oral administration of (S)-55 at 100 mg/kg resulted in a 69% reduction of plasma Aβ40 at 8 h in a Tg2576 mouse ($p < 0.001$).

1.3. Design, Synthesis and SAR of Potent Statine-Based BACE-1 Inhibitors

Marcus Bäck *et al.* [69] presented several BACE-1 inhibitors with low nanomolar level activities, encompassing a statine-based core structure with phenyloxymethyl- and benzyloxymethyl residues in the P1 position. The authors introduced a novel P1 modification to allow the facile exploration of the S1 binding pocket of BACE-1, delivering highly promising inhibitors. Recently reported disclose inhibitors comprising the central cores II with a phenyloxymethyl P1 group for which the norstatine thio analog. In addition, they present inhibitors encompassing the extended central core III with a benzyloxymethyl residue in the P1 position (Fig. **4**). Flexible synthesis provided compounds exhibiting IC_{50} values in the low nanomolar range.

1.4. Macrocyclic Tertiary Carbinamine BACE-1 Inhibitors

In this paper, Stacey R. Lindsley *et al.* [70] described the design and synthesis of tertiary carbinamine macrocyclic inhibitors of the β-secretase (BACE-1) enzyme. These macrocyclic inhibitors, some of which incorporate novel P2 substituents, display a 2- to 100-fold increase in potency relative to the previously described acyclic analogues while affording greater stability (Fig. **5**).

1.5. Conformationally Biased P3 Amide Replacements of β-Secretase Inhibitors

Other authors, such as Shawn J. Stachel *et al.* [71] synthesized and evaluated a series of conformationally biased P3 amide replacements based on an isophthalamide lead structure. These studies resulted in the identification of the β-secretase inhibitor 7m which has an *in vitro* IC50 = 35 nM. This synthesis and the biological activities of these compounds are described in this paper (Fig. **6**).

2. QSAR and Theoretical Studies for β-Secretase Inhibitors

In this section we update the contents presented in our recent review published in Current Drugs Metabolism [72]. The high number of possible candidates to β-secretase inhibitors creates the necessity of Quantitative Structure-Activity Relationship models in order to guide the β-secretase inhibitor synthesis. In this work, we reviewed different computational studies for a very large and heterogeneous series of β-secretase. First, we reviewed QSAR studies with conceptual parameters. Next, using the method of the regression analysis and QSAR studies in order to understand the essential structural requirement for receptor binding. (la frase no tiene predicado) Next, we reviewed 3D QSAR, CoMFA and CoMSIA with different compounds to find out the structural requirements for β-secretase inhibitors.

2.1. Models of Novel Pyridinium-Based Potent β-Secretase Inhibitory Leads

In the paper carried out by Afaf Al-Nadaf *et al.* [73], the authors explored the pharmacophoric space of 129 known BACE inhibitors have potential as anti-Alzheimer's disease treatments. The QSAR analysis was employed to select the optimal combination of pharmacophoric models and 2D physicochemical descriptors capable of explaining bioactivity variation ($r^2 = 0.88$, $F = 60.48$, r^2 LOO = 0.85, r^2 PRESS against 25 external test inhibitors = 0.71). They were obliged to use ligand efficiency as response variable because the logarithmic transformation of bioactivities failed to access self-consistent QSAR models. The authors created three pharmacophoric models emerged in the successful QSAR equation, suggesting at least three binding modes accessible to ligands within BACE binding pocket. The QSAR equation and pharmacophoric models were validated through ROC curves (Table **1**), and were employed to guide synthesis of novel pyridinium-based BACE inhibitors.

2.2. CoMFA & CoMSIA of Hydroxyethylamine Derivatives as BACE-1 Inhibitors

Ashish Pandey *et al.* [74] developed three-dimensional quantitative structure–activity relationship (3D-QSAR) models based on comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA), on a series of 43 hydroxyethylamine derivatives, acting as potent inhibitors of *β*-site amyloid precursor protein (APP) cleavage enzyme (BACE-1). They used a crystal structure of the BACE-1 enzyme (PDB ID: 2HM1) with one of the most active compound presented in this paper was available, and we assumed it to be the bioactive conformation of the studied series, for a 3D-QSAR analysis. A statistically significant 3D-QSAR model was established on a training set of 34 compounds, which were validated by a test set of 9 compounds. For the best CoMFA model, the statistics are $r^2 = 0.998$, $r^2_{cv} = 0.810$, $n = 34$ for the training set and $r^2_{pred} = 0.934$, $n = 9$ for the test set. For

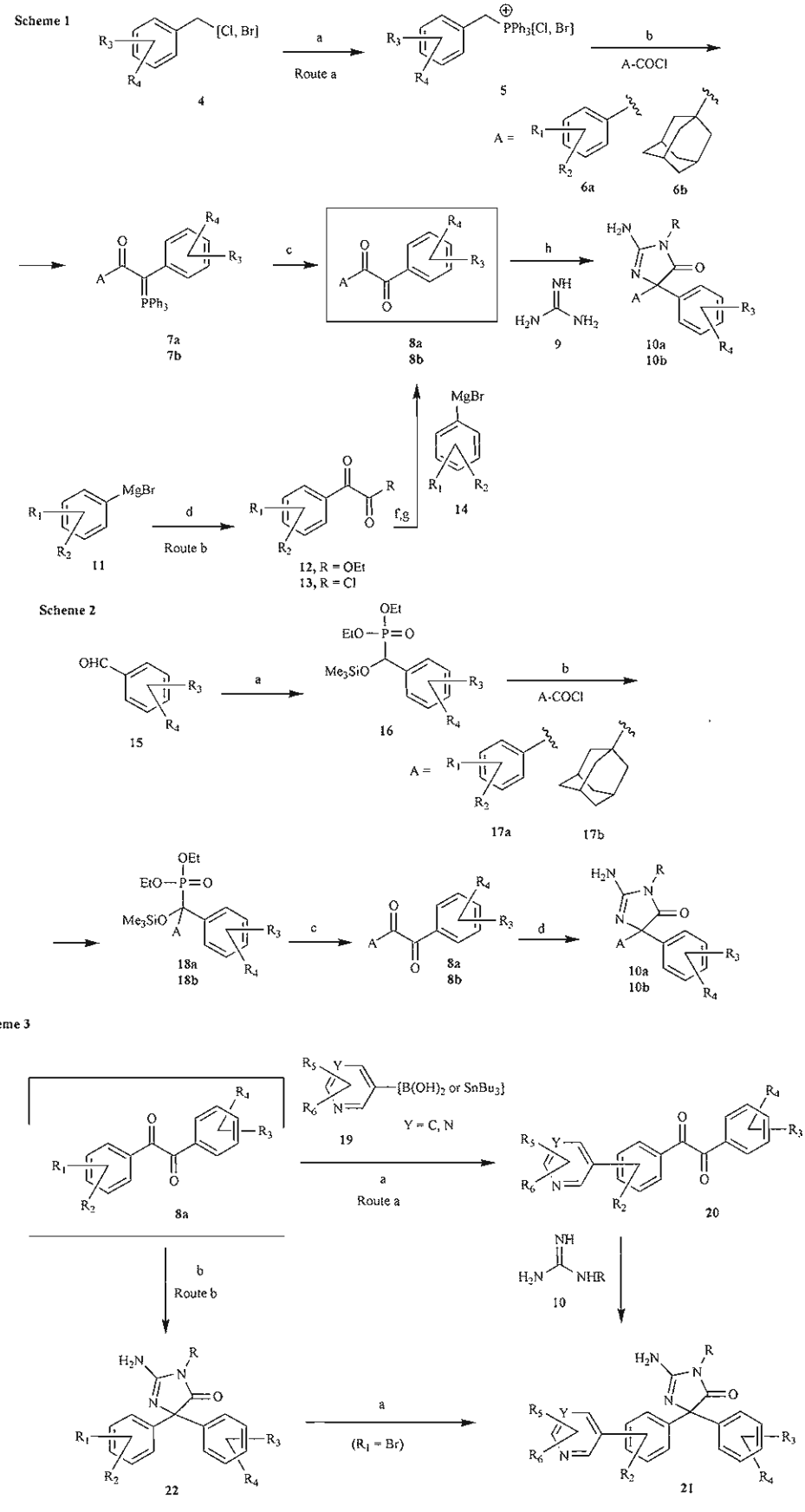
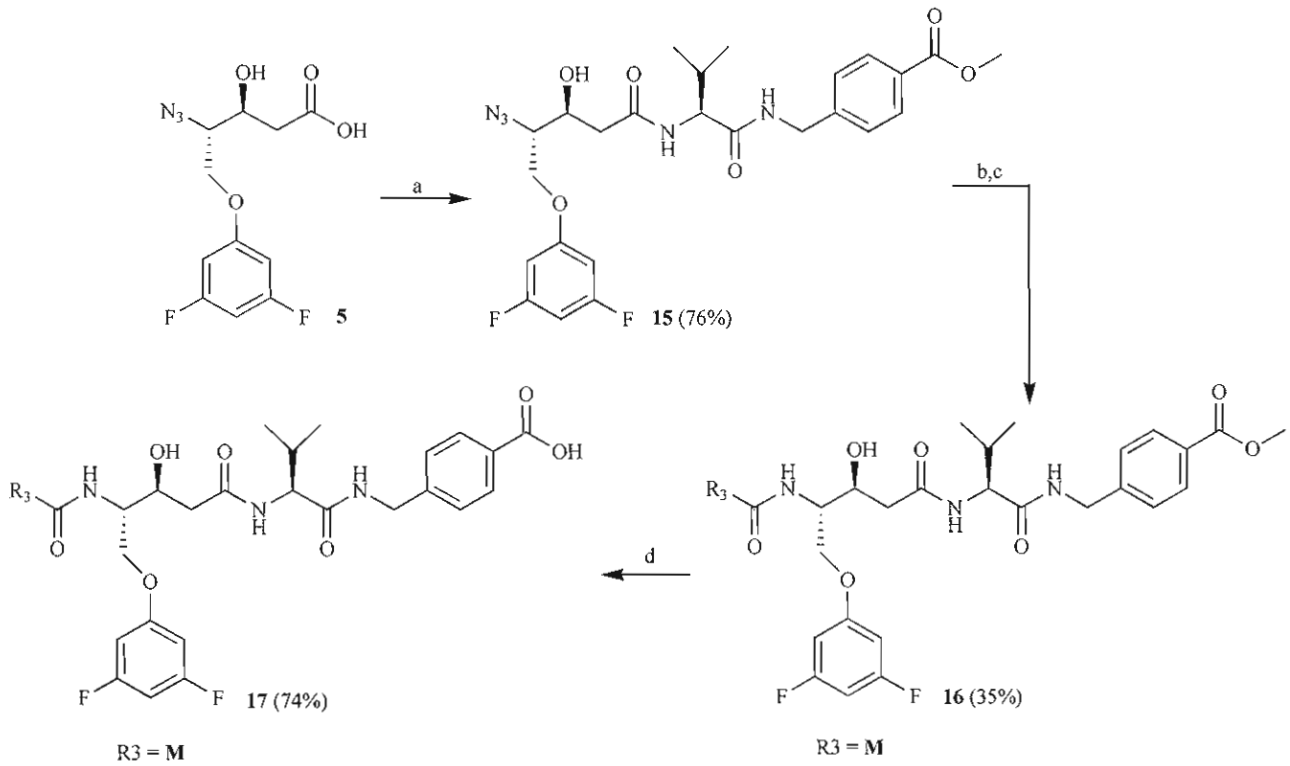
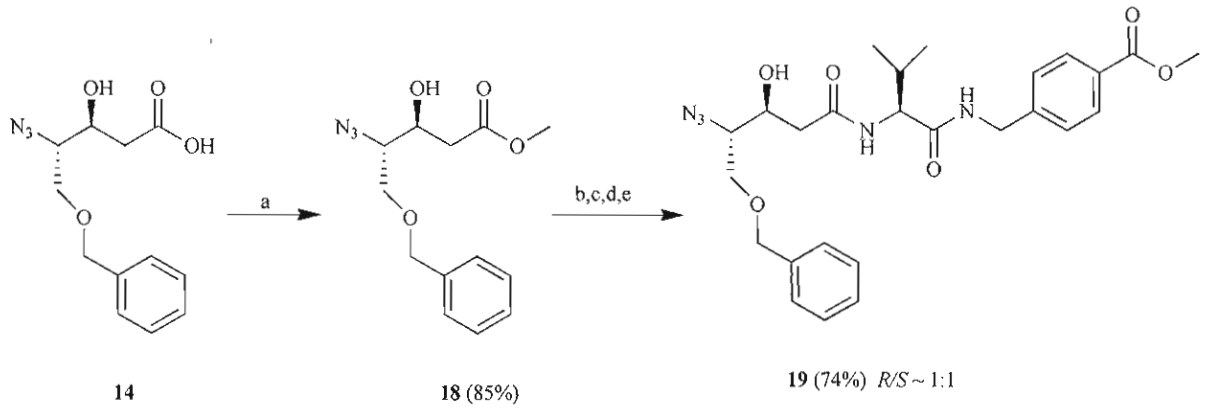


Fig. (3). Synthesis of Aminohydantoins β-secretase inhibitors.

Scheme 1



Scheme 2

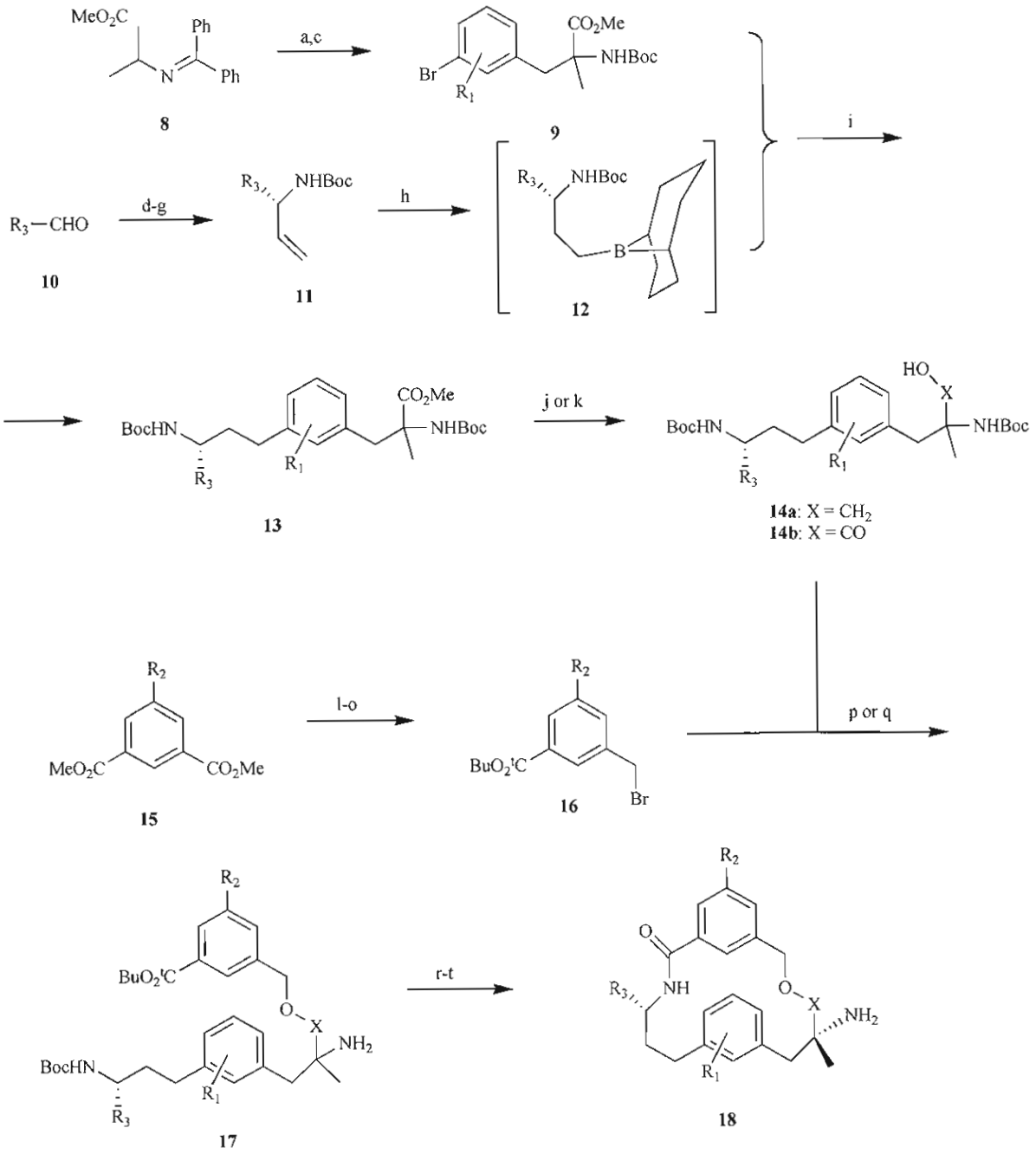


Scheme 1: Reagents and conditions: (a) A, DIPEA, HATU, DMF; (b) PPh_3, MeOH, H_2O; (c) M, DIPEA, HATU, DMF; (d) 1 M LiOH, THF/MeOH/H_2O 2:1:1.
Scheme 2: Reagents and conditions: (a) $SOCl_2$, MeOH; (b) Dess-Martin periodinane, DCM; (c) $NaBH_4$, MeOH, 15°C; (d) LiOH, dioxane/H_2O 1:1; (e) A, DIPEA, HATU, DMF.

Fig. (4). Synthesis of P1 phenoxy and benzyloxy residues.

the best CoMSIA model (combined steric, electrostatic, hydrophobic, and hydrogen bond donor fields), the statistics are $r^2 = 0.978$, $r^2_{cv} = 0.754$, $n = 34$ for the training set and $r^2_{pred} = 0.750$, $n = 9$ for the test set (Table **2**). The resulting contour maps, produced by the best CoMFA and CoMSIA models, were used to identify the structural features relevant to the biological activity in series of analogues. The data generated from this study will further help to design novel, potent, and selective BACE-1 inhibitors.

Scheme 1



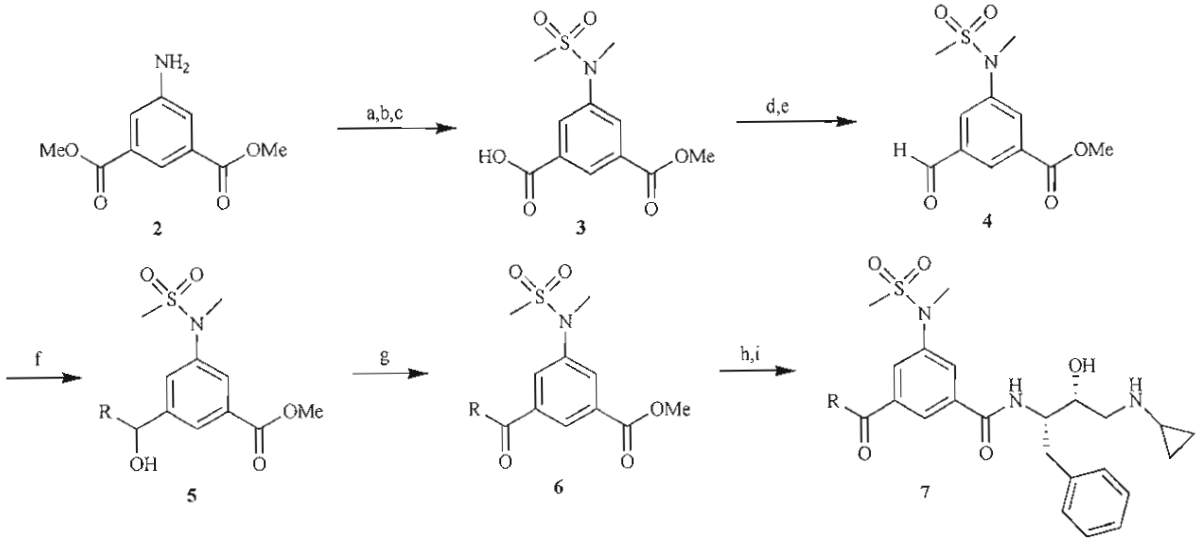
Reagents and conditions: (a) base, $ArCH_2Br$; (b) HCl, MeOH; (c) Boc_2O, DIEA; (d) Ellman sulfinimine; (e) vinyl Grignard, THF; (f) HCl, MeOH; (g) Boc_2O, Hunig's base; (h) 9-BBN, THF; (i) $Pd(PPh_3)_4$, 3 N NaOH, toluene, 85°C; (j) for 14a: $LiBH_4$,THF; (k) for 14b: 1 N LiOH, THF; (l) 1 equiv NaOH, MeOH; (m) CDI, t-BuOH; (n) $LiBH_4$,THF; (o) CBr_4, PPh_3, DCM; (p) for 14a: AgOTf, 2,6-di-t- Bu-pyr resin, DCE; (q) for 14b: Cs_2CO_3, DMF; (r) HCl, DCM; (s) BOP, DIEA, DMF; (t) chiral preparative HPLC; for R_2 = Br: Boc_2O; flash chromatography; R_2-M, Pd(0); HCl, DCM.

Fig. (5). Synthesis macrocyclic tertiary carbinamine BACE-1 inhibitors.

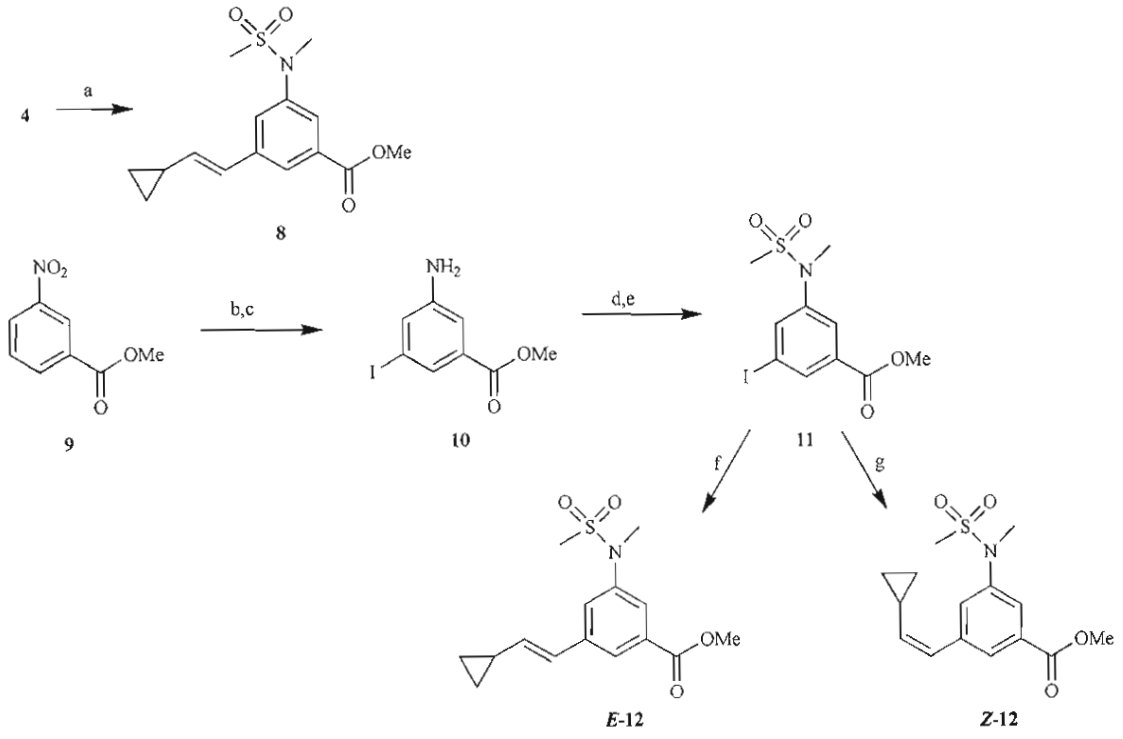
2.3. Virtual Screening and Protonation States at Asp32 and Asp228

György M. Keseru *et al.* [75] performed a comparative virtual screen for β-secretase (BACE-1) inhibitors using different docking methods (FlexX and FlexX-Pharm), scoring functions (Dock, Gold, Chem, PMF, FlexX), protonation states (default and calculated), and protein conformations (apo and ligand bound). The apo- and ligand-bound conformations of BACE-1 were both found to be

Scheme 1



Scheme 2



Scheme 1 Reagents: (a) MsCl, pyridine, CH_2Cl_2, 75%; (b) NaH, MeI, DMF, 98%; (c) 0.1 N LiOH, THF/H_2O, 67%; (d) EDC, HOBT, N,O-dimethylhydroxylamine HCl, $NaHCO_3$, 81%; (e) DIBAL-H, 94%; (f) RMgX; (g) MnO_2, CH_2Cl_2; (h) 1N NaOH, THF/MeOH/H_2O (i) BOP, hydroxyethyldiamine, diisopropylethylamine.

Scheme 2. Reagents: (a) Cyclopropylmethyl triphenylphosphonium bromide, n-BuLi, 46%; (b) TfOH, NIS, 37%; (c) $SnCl_2$ Æ H_2O, THF/ EtOH, 96%, (d) MsCl, pyridine, CH_2Cl_2, 85%; (e) NaH, MeI, DMF, 98%; (f) cyclopropylacetylene, 9-BBN, $PdCl_2$ Æ (dppf) CH_2Cl_2, $AsPh_3$, Cs_2CO_3, 73%; (g) cyclopropylacetylene, $InCl_3$, DIBAL-H, Et_3B, $Pd(dba)_2$ Æ $CHCl_3$, P(furyl)$_3$, 93%.

Fig. (6). Synthesis of biased P3 amide replacements of β-secretase inhibitors.

suitable for virtual screening. Assigning calculated protonation states to catalytic Asp32 and Asp228 residues resulted in significant improvement of enrichment factors as calculated at 1% of the ranked database. The authors used 1FKN to obtain no enrichment by FlexX/D-Score that was improved to ligand when considering calculated protonation states. They also showed that combining calculated protonation states with pharmacophore constraints using FlexX-Pharm/D-Score improved enrichment further to ligand. Enrichments reported in this study suggested that our screening protocol would be effective in the virtual screening of large compound libraries for BACE-1 inhibitors.

Table 1. ROC$_a$ Performances of QSAR-Selected Pharmacophores as 3D Search Queries

Pharmacophore	ROC$_a$/AUC$_b$	ACC$_c$	SPC$_d$	TPR$_e$	FNR$_f$
Hypo10/10	0.982	0.961	0.988	0.28	0.011345
Hypo6/18	0.981	0.961	0.975	0.6	0.024311
Hypo1/21	0.738	0.961	0.9611	0.96	0.038898

[a]ROC: receiver operating characteristic, [b]AUC: area under the curve, [c] ACC: overall accuracy, [d]SPC: overall specificity, [e] TPR: overall true positive rate, [f]FNR. overall false negative rate.

Table 2. PLS Summary of CoMFA and CoMSIA Results

Statistical Parameters	CoMFA (S E)	CoMSIA (S EHD)
Number of molecules in training set	34	34
Number of molecules in test set	9	9
r^2 cv	0.810	0.754
NOC	7	4
SEE	0.063	0.204
r^2	0.998	0.978
F-test	2009.08	324.673
r^2_{bs}	0.999	0.989
SD_{bs}	0.001	0.006
r^2_{pred}	0.934	0.750
Percentage of field contributions		
S	47.4	24.8
E	52.6	34.0
H	–	26.3
D	–	14.9

Abbreviations: S (steric field), E (electrostatic field), H (hydrophobic field), D (hydrogen bond donor field) r^2_{cv} =Cross-validated correlation coefficient by PLS LOO method, NOC=Optimum number of components as determined by PLS LOO cross-validation study, SEE=Standard error of estimate, r^2 =Conventional correlation coefficient, r^2_{bs} =Correlation coefficient after 100 runs of boot strapping, SD_{bs} =Standard deviation from 100 runs of bootstrapping, r^2_{pred} =Predictive correlation coefficient.

2.4. Docking Scoring Function Based on 2D-Descriptors

In this paper, Csaba Hetényi [76] showed a key step in the molecular engineering of such potent lead compounds, which consists of the prediction of the energetics of their binding to the macromolecular targets. Although sophisticated experimental and *in silico* methods are available to solve this issue, the structure-based calculation of the binding free energies of large, flexible ligands to proteins is problematic. In this study, a fast and accurate calculation strategy was presented, followed by a modification of the scoring function of the popular docking program package AutoDock and the involvement of ligand-based two-dimensional descriptors. Quantitative structure-activity relationships with good predictive power were developed. The best results of this paper were shown in Table 3. Cross-validation tests and verifications were performed on the basis of experimental binding data of biologically important systems. The capabilities and limitations of the ligand-based descriptors were analyzed. According to the authors, the application of these results in the early phase of lead design would contribute to precise predictions, correct selections, and consequently a higher success rate of rational drug discovery.

2.5. Induced-Fit Docking of Peptidic and Pseudo-Peptidic BACE-1 Inhibitors

Inhibition of β-secretase (BACE-1) has recently been investigated as a promising therapeutic approach in the treatment of Alzheimer's disease, and a growing number of BACE-1 inhibitors and crystal structures of BACE-1/inhibitor complexes have been reported. Nicolas Moitessier *et al.* [77] reported a predictive computational method and its application to potential BACE-1 inhibitors. Using a training set of 50 known highly flexible inhibitors, they developed a docking method that accounted for the flexibility of both the protein and the inhibitors. Protein flexibility was considered due to the use of a specifically designed genetic algorithm. In this paper, the authors developed a scoring function consisting of force field evaluation of the inhibitor/protein interactions and two additional terms for hydrogen bonding and entropy change upon binding. Discarding three outliers from the training set, the protocol was found to perform well with an rmsd of 1.19 kcal/mol an/and r^2 value of 0.789. The evaluation of the predictive power was carried out by virtual screening of 80 synthetic compounds. The significant enrichment at the top of the ranking list in active compounds demonstrated the ability of the docking and scoring protocol to rank the compounds according to their activities.

3. New Method for the Study of New β-Secretase Inhibitors

3.1. Preface to New QSAR Study β-Secretase Inhibitors

Alzheimer's disease (AD) [78] is a serious and degenerative disorder that causes agradual loss of neurons, and in spite of the efforts performed by the big pharmaceutical companies worldwide, the origin of this pathology is still not very clear. We can see in this paper that the development of theoretical and QSAR models to study β-secretase inhibitors is not completely achieved so far, and most of these works present docking studies. Considering this situation we need to develop QSAR models with β-secretase inhibitors. In this sense, quantitative structure-activity relationships (QSAR) could play an important role in studying these β-secretase inhibitors; QSARs can be used as predictive tools for the development of molecules [79, 80]. Moreover, computer-aided drug design techniques based on Quantitative Structure-Activity Relationships (QSAR) could play an important role in drug discovery programs. The QSAR approach involves the development of models that relate the structure of drugs with their biological activity against different targets [81, 82]. In principle, there are currently more than 1600 molecular descriptors that may be generalized and used to solve the problem outlined above [83]. Numerous different molecular descriptors have been

Table 3. The Results Produced by the Best CoMFA and CoMSIA Models

QSAR	Descriptor (Di)							
		Coefficient (Ri)	Error of Coeff.	t-Value	R^2	R^2_{cv}	s^2	F-Value
A	1	¢GTH	3.1216 × 10-1	2.4686 × 10-2	0.799	0.774	1.05	93.36
	2	RPCGEN	3.2582 × 101	6.9963				
		constant	-4.1980	6.9930 × 10-1				
B	1	¢GTH	2.7077 × 10-1	2.2926 × 10-2	0.859	0.838	0.76	93.17
	2	RPCGEN	5.7129 × 101	8.1307				
	3	J	-6.2410 × 10-1	1.4148 × 10-1				
		constant	-4.6864	6.0281 × 10-1				

Standard deviations (s^2), squares of the correlation coefficients (R^2), and leave-one-out cross-validated correlation coefficients (R^2_{cv}) of the regressions are tabulated.

reported to encode chemical structures in QSAR studies. Furthermore, there are multiple chemometric approaches that can, in principle, be selected for this step. Multiple linear regression (MLR) or linear discriminant analysis (LDA) [84], can be used to relate molecular structure (represented by molecular descriptors) with biological properties. In this article, we developed QSAR models for β-secretase inhibitors, LDA was constructed from more than 50000 cases with more than 3000 different molecules inhibitors of β-secretase obtained from ChEMBL database http://www.ebi.ac.uk/chembldb/index.php; in total we used more than 15000 different molecules to develop the QSAR models. We used spectral moments (molecular descriptors) calculated with the ModesLab software [85].

3.2. Methods

3.2.1. Linear Classifier

A database from ChEMBL database [86] containing assayed β-secretase inhibitors was used (Table **SM** from the Supplementary Material, requested to the author). The ModesLab software [85] was employed and provided spectral moments (descriptors) [87-91]. The QSAR model was constructed with the multivariate regression technique, the LDA, employing the Forward stepwise method for the selection of variables. All statistical analyses and data exploration were carried out in STATISTICA 6.0 [92]. In the present work, the independent data test is used by splitting the data randomly in a training series employed for a model construction and a cross-validation (CV) one. The general formula of the QSAR classification function is the following:

$$\beta - \text{sec}\,retase_{score} = \sum W_m \cdot {}^m 2D_i + W_0 \qquad (1)$$

where β-secretase$_{score}$ is the continuous and dimensionless score value for the β-secretase$_{score}$ /non-β-secretase$_{score}$ classification that gives relatively higher values to molecules with more probability to act as β-secretase$_{score}$, ${}^m 2D_i$ are the $2Ds$ of type m, W_m is the coefficient (weights) of these indices in the QSAR model and W_0 is the independent term. The reported statistical parameters of the QSAR model are the following: N, χ^2, F, and p-level as well as Sensitivity, Specificity and Accuracy for both training and CV. N is the number of molecules used to train the model, λ is Wilks statistic parameter, is Chi-square and p-level is the probability of error.

3.2.2. Data Set

The data set used in this paper was obtained from the ChEMBL database. It has more than 20000 cases and more than 3000 different compounds, inhibitors of β-secretase. In total we used more than 15000 different molecules to develop the QSAR models obtained in ChEMBL. This is a database of bioactive drug-like small molecules, it contains 2D structures, calculated properties (e.g. logP, Molecular Weight, Lipinski Parameters, etc.) and abstracted bioactivities (e.g. binding constants, pharmacology and ADMET data). ChEMBL normalises the bioactivities into a uniform set of end-points and units where possible, and also tags the links between a molecular target and a published assay with a set of varying confidence levels. The data are abstracted and curated from the primary scientific literature, and cover a significant fraction of the SAR and discovery of modern drugs. The codes and activity for all compounds, as well as the references used to collect them are depicted in **SM1** of the supplementary material file (requested to the author).

3.3. Results and Discussion

The present is a QSAR model for the probability of binding organic compounds to β-secretase receptor based only on the molecular connectivity of the drug and the protein receptor. Using this model we can predict the different relationships between the drug-protein connectivity same, that is, physicochemical property [93]. This work introduces a single linear QSAR equation model to classify drugs with β-secretase receptor. The best model found was:

$$\begin{aligned}\beta - \text{sec}_{pred} = {} & 2.78 \cdot \mu_0 - 0.21 \cdot \mu_2 + 15.47 \cdot \mu_2 - 5.3 \cdot \mu_4 - 8.96 \\ & \cdot \mu_5 + 4.13 \cdot \mu_6 + 1.71 \cdot \mu_7 - 1.11 \cdot \mu_8 + 0.09 \\ & \cdot \mu_{10} - 0.02 \cdot \mu_{11} + 0.002 \cdot \mu_{12} - 0.0002 \\ & \cdot \mu_{13} + 2.6x10^{-6} \cdot \mu_{15} - 7.9196 \\ N = 22323 \quad & \chi 2 = 11856.74 \quad \lambda = 0.451 \quad p < 0.001\end{aligned} \qquad (2)$$

The nomenclature used in the descriptors of the equation is the same as the one establishing/established by the Dragon software, where N is the number of compounds used for training, λ is the Wilks's statistic parameter, χ^2 is the Chi-

square and p is the level of error. This model, with 13 variables, classified correctly 1803 out of 2104 active (Sensitivity of 85.69%) and 12168 out of 12825 non-active (Specificity of 94.88%). Overall training Accuracy was 93.58%. The validation of the model was carried out by means of external predicting series. The model classified correctly 921 out of 1053 active (87.46%) and 6043 out of 6341 non-active (95.3%) in validation series. Accuracy for validation series (predictability) was 94.18% (6964 out of 7394 DTPs). These results (Table **4**) indicate that we have developed an accurate model according to previous reports on the use of LDA in QSAR [94, 95].

Table 4. Results of the New LDA Classification Model

Parameters	%	Class	Active	Non-Active
Analysis				
Sensitivity	94.88	Non-active	12168	657
Specificity	85.69	Active	301	1803
Accuracy	93.58	Total		
Validation				
Sensitivity	95.30	Non-active	6043	298
Specificity	87.46	Active	132	921
Accuracy	94.18	Total		

CONCLUSIONS

The functions of β-secretase and its implication in Alzheimer's disease have triggered an active search for potent and selective β-secretase inhibitors. Nowadays theoretical studies such as QSAR models have become a very useful tool in this context, to substantially reduce time and resource-consuming experiments. In this paper we can see that the development of theoretical and QSAR models to study β-secretase inhibitors is not completely achieved so far, and most of these works present docking studies. Considering this situation, we need to develop QSAR models with β-secretase inhibitors. In this sense, QSAR could play an important role in studying these β-secretase inhibitors. Moreover, QSARs can be used as predictive tools for the development of molecules. In this work we have developed a new LDA model using the ModesLab descriptors, based on a large database consisting of about 15,000 different drugs, obtained from the ChemBL server.

ACKNOWLEDGEMENTS

F. Prado-Prado thanks financial support for the research position at the University of Santiago de Compostela from *Angeles* Alvariño, Xunta de Galicia. All the authors acknowledge support from the Project 07CSA008203PR.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's web site along with the published article.

REFERENCES

[1] Goate, A.; Chartier-Harlin, M.C.; Mullan, M.; Brown, J.; Crawford, F.; Fidani, L.; Giuffra, L.; Haynes, A.; Irving, N.; James, L.; *et al.* Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature*, **1991**, *349*, 704-706.

[2] Schellenberg, G.D.; Bird, T.D.; Wijsman, E.M.; Orr, H.T.; Anderson, L.; Nemens, E.; White, J.A.; Bonnycastle, L.; Weber, J.L.; Alonso, M.E.; *et al.* Genetic linkage evidence for a familial Alzheimer's disease locus on chromosome 14. *Science*, **1992**, *258*, 668-671.

[3] Golde, T.E.; Dickson, D.; Hutton, M. Filling the gaps in the abeta cascade hypothesis of Alzheimer's disease. *Curr. Alzheimer Res.*, **2006**, *3*, 421-430.

[4] Hutton, M.; Perez-Tur, J.; Hardy, J. Genetics of Alzheimer's disease. *Essays Biochem.*, **1998**, *33*, 117-131.

[5] Younkin, S.G. The role of A beta 42 in Alzheimer's disease. *J. Physiol. Paris*, **1998**, *92*, 289-292.

[6] Sisodia, S.S. Alzheimer's disease: perspectives for the new millennium. *J. Clin. Invest.*, **1999**, *104*, 1169-1170.

[7] Rovelet-Lecrux, A.; Hannequin, D.; Raux, G.; Le Meur, N.; Laquerriere, A.; Vital, A.; Dumanchin, C.; Feuillette, S.; Brice, A.; Vercelletto, M.; Dubas, F.; Frebourg, T.; Campion, D. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat. Genet.*, **2006**, *38*, 24-26.

[8] Ohno, M.; Sametsky, E.A.; Younkin, L.H.; Oakley, H.; Younkin, S.G.; Citron, M.; Vassar, R.; Disterhoft, J.F. BACE1 deficiency rescues memory deficits and cholinergic dysfunction in a mouse model of Alzheimer's disease. *Neuron*, **2004**, *41*, 27-33.

[9] Ohno, M.; Cole, S.L.; Yasvoina, M.; Zhao, J.; Citron, M.; Berry, R.; Disterhoft, J.F.; Vassar, R. BACE1 gene deletion prevents neuron loss and memory deficits in 5XFAD APP/PS1 transgenic mice. *Neurobiol. Dis.*, **2007**, *26*, 134-145.

[10] Laird, F.M.; Cai, H.; Savonenko, A.V.; Farah, M.H.; He, K.; Melnikova, T.; Wen, H.; Chiang, H.C.; Xu, G.; Koliatsos, V.E.; Borchelt, D.R.; Price, D.L.; Lee, H.K.; Wong, P.C. BACE1, a major determinant of selective vulnerability of the brain to amyloid-beta amyloidogenesis, is essential for cognitive, emotional, and synaptic functions. *J. Neurosci.*, **2005**, *25*, 11693-11709.

[11] Selkoe, D.J. Alzheimer's disease: genes, proteins, and therapy. *Physiol. Rev.*, **2001**, *81*, 741-766.

[12] Vassar, R. BACE1: the beta-secretase enzyme in Alzheimer's disease. *J. Mol. Neurosci.*, **2004**, *23*, 105-114.

[13] Hussain, I.; Powell, D.; Howlett, D.R.; Tew, D.G.; Meek, T.D.; Chapman, C.; Gloger, I.S.; Murphy, K.E.; Southan, C.D.; Ryan, D.M.; Smith, T.S.; Simmons, D.L.; Walsh, F.S.; Dingwall, C.; Christie, G. Identification of a novel aspartic protease (Asp 2) as beta-secretase. *Mol. Cell Neurosci.*, **1999**, *14*, 419-427.

[14] Sinha, S.; Anderson, J.P.; Barbour, R.; Basi, G.S.; Caccavello, R.; Davis, D.; Doan, M.; Dovey, H.F.; Frigon, N.; Hong, J.; Jacobson-Croak, K.; Jewett, N.; Keim, P.; Knops, J.; Lieberburg, I.; Power, M.; Tan, H.; Tatsuno, G.; Tung, J.; Schenk, D.; Seubert, P.; Suomensaari, S.M.; Wang, S.; Walker, D.; Zhao, J.; McConlogue, L.; John, V. Purification and cloning of amyloid precursor protein beta-secretase from human brain. *Nature*, **1999**, *402*, 537-540.

[15] Vassar, R.; Bennett, B.D.; Babu-Khan, S.; Kahn, S.; Mendiaz, E.A.; Denis, P.; Teplow, D.B.; Ross, S.; Amarante, P.; Loeloff, R.; Luo, Y.; Fisher, S.; Fuller, J.; Edenson, S.; Lile, J.; Jarosinski, M.A.; Biere, A.L.; Curran, E.; Burgess, T.; Louis, J.C.; Collins, F.; Treanor, J.; Rogers, G.; Citron, M. Beta-secretase cleavage of Alzheimer's amyloid precursor protein by the transmembrane aspartic protease BACE. *Science*, **1999**, *286*, 735-741.

[16] Yan, R.; Bienkowski, M.J.; Shuck, M.E.; Miao, H.; Tory, M.C.; Pauley, A.M.; Brashier, J.R.; Stratman, N.C.; Mathews, W.R.; Buhl, A.E.; Carter, D.B.; Tomasselli, A.G.; Parodi, L.A.; Heinrikson, R.L.; Gurney, M.E. Membrane-anchored aspartyl protease with Alzheimer's disease beta-secretase activity. *Nature*, **1999**, *402*, 533-537.

[17] Haass, C.; Schlossmacher, M.G.; Hung, A.Y.; Vigo-Pelfrey, C.; Mellon, A.; Ostaszewski, B.L.; Lieberburg, I.; Koo, E.H.; Schenk, D.; Teplow, D.B.; *et al.* Amyloid beta-peptide is produced by cultured cells during normal metabolism. *Nature*, **1992**, *359*, 322-325.

[18] Seubert, P.; Oltersdorf, T.; Lee, M.G.; Barbour, R.; Blomquist, C.; Davis, D.L.; Bryant, K.; Fritz, L.C.; Galasko, D.; Thal, L.J.; *et al.* Secretion of beta-amyloid precursor protein cleaved at the amino terminus of the beta-amyloid peptide. *Nature*, **1993**, *361*, 260-263.

[19] Zhao, J.; Paganini, L.; Mucke, L.; Gordon, M.; Refolo, L.; Carman, M.; Sinha, S.; Oltersdorf, T.; Lieberburg, I.; McConlogue, L. Beta-secretase

processing of the beta-amyloid precursor protein in transgenic mice is efficient in neurons but inefficient in astrocytes. *J. Biol. Chem.*, **1996**, *271*, 31407-31411.

[20] Citron, M.; Teplow, D.B.; Selkoe, D.J. Generation of amyloid beta protein from its precursor is sequence specific. *Neuron*, **1995**, *14*, 661-670.

[21] Santana, L.; Uriarte, E.; González-Díaz, H.; Zagotto, G.; Soto-Otero, R.; Mendez-Alvarez, E. A QSAR model for *in silico* screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins. *J. Med. Chem.*, **2006**, *49*, 1149-1156.

[22] Marrero-Ponce, Y. Linear indices of the "molecular pseudograph's atom adjacency matrix": definition, significance-interpretation, and application to QSAR analysis of flavone derivatives as HIV-1 integrase inhibitors. *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 2010-2026.

[23] Vilar, S.; Santana, L.; Uriarte, E. Probabilistic neural network model for the *in silico* evaluation of anti-HIV activity and mechanism of action. *J. Med. Chem.*, **2006**, *49*, 1118-1124.

[24] Marrero-Ponce, Y.; Khan, M.T.; Casanola Martin, G.M.; Ather, A.; Sultankhodzhaev, M.N.; Torrens, F.; Rotondo, R. Prediction of tyrosinase inhibition activity using atom-based bilinear indices. *ChemMedChem*, **2007**, *2*, 449-478.

[25] Casanola-Martin, G.M.; Marrero-Ponce, Y.; Khan, M.T.; Ather, A.; Sultan, S.; Torrens, F.; Rotondo, R. TOMOCOMD-CARDD descriptors-based virtual screening of tyrosinase inhibitors: evaluation of different classification model combinations using bond-based linear indices. *Bioorg. Med. Chem.*, **2007**, *15*, 1483-1503.

[26] Casanola-Martin, G.M.; Marrero-Ponce, Y.; Khan, M.T.; Ather, A.; Khan, K.M.; Torrens, F.; Rotondo, R. Dragon method for finding novel tyrosinase inhibitors: Biosilico identification and experimental *in vitro* assays. *Eur. J. Med. Chem.*, **2007**, *42*, 1370-1381.

[27] Nunez, M.B.; Maguna, F.P.; Okulik, N.B.; Castro, E.A. QSAR modeling of the MAO inhibitory activity of xanthones derivatives. *Bioorg. Med. Chem. Lett.*, **2004**, *14*, 5611-5617.

[28] Terada, M.; Inaba, M.; Yano, Y.; Hasuma, T.; Nishizawa, Y.; Morii, H.; Otani, S. Growth-inhibitory effect of a high glucose concentration on osteoblast-like cells. *Bone*, **1998**, *22*, 17-23.

[29] González-Díaz, H.; González-Díaz, Y.; Santana, L.; Ubeira, F.M.; Uriarte, E. Proteomics, networks and connectivity indices. *Proteomics*, **2008**, *8*, 750-778.

[30] Zhao, C.J.; Dai, Q.Y. Recent advances in study of antinociceptive conotoxins. *Yao Xue Xue Bao*, **2009**, *44*, 561-565.

[31] Jacob, R.B.; McDougal, O.M. The M-superfamily of conotoxins: a review. *Cell. Mol. Life Sci.*, **2010**, *67*, 17-27.

[32] Giuliani, A.; Di Paola, L.; Setola, R. Proteins as networks: a mesoscopic approach using haemoglobin molecule as case study. *Curr. Proteomics*, **2009**, *6*, 235-245.

[33] Vilar, S.; Gonzalez-Diaz, H.; Santana, L.; Uriarte, E. A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer. *J. Theoret. Biol.*, **2009**, *261*, 449-458.

[34] Concu, R.; Dea-Ayuela, M.A.; Perez-Montoto, L.G.; Prado-Prado, F.J.; Uriarte, E.; Bolas-Fernandez, F.; Podda, G.; Pazos, A.; Munteanu, C.R.; Ubeira, F.M.; Gonzalez-Diaz, H. 3D entropy and moments prediction of enzyme classes and experimental-theoretic study of peptide fingerprints in Leishmania parasites. *Biochimica et Biophysica Acta*, **2009**, *1794*, 1784-1794.

[35] Torrens, F.; Castellano, G. Topological charge-transfer indices: from small molecules to proteins. *Curr. Proteomics*, **2009**, 204-213.

[36] Vázquez, J.M.; Aguiar, V.; Seoane, J.A.; Freire, A.; Serantes, J.A.; Dorado, J.; Pazos, A.; Munteanu, C.R. Star graphs of protein sequences and proteome mass spectra in cancer prediction. *Curr. Proteomics*, **2009**, *6*, 275-288.

[37] Gonzalez-Diaz, H. Quantitative studies on Structure-Activity and Structure-Property Relationships (QSAR/QSPR). *Curr. Top. Med. Chem.*, **2008**, *8*, 1554.

[38] Ivanciuc, O. Weka machine learning for predicting the phospholipidosis inducing potential. *Curr. Top. Med. Chem.*, **2008**, *8*, 1691-1709.

[39] Gonzalez-Diaz, H.; Prado-Prado, F.; Ubeira, F.M. Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. *Curr. Top. Med. Chem.*, **2008**, *8*, 1676-1690.

[40] Duardo-Sanchez, A.; Patlewicz, G.; Lopez-Diaz, A. Current topics on software use in medicinal chemistry: intellectual property, taxes, and regulatory issues. *Curr. Top. Med. Chem.*, **2008**, *8*, 1666-1675.

[41] Wang, J.F.; Wei, D.Q.; Chou, K.C. Drug candidates from traditional chinese medicines. *Curr. Top. Med. Chem.*, **2008**, *8*, 1656-1665.

[42] Helguera, A.M.; Combes, R.D.; Gonzalez, M.P.; Cordeiro, M.N. Applications of 2D descriptors in drug design: a DRAGON tale. *Curr. Top. Med. Chem.*, **2008**, *8*, 1628-1655.

[43] Gonzalez, M.P.; Teran, C.; Saiz-Urra, L.; Teijeira, M. Variable selection methods in QSAR: an overview. *Curr. Top. Med. Chem.*, **2008**, *8*, 1606-1627.

[44] Caballero, J.; Fernandez, M. Artificial neural networks from MATLAB in medicinal chemistry. Bayesian-regularized genetic neural networks (BRGNN): application to the prediction of the antagonistic activity against human platelet thrombin receptor (PAR-1). *Curr. Top. Med. Chem.*, **2008**, *8*, 1580-1605.

[45] Wang, J.F.; Wei, D.Q.; Chou, K.C. Pharmacogenomics and personalized use of drugs. *Curr. Top. Med. Chem.*, **2008**, *8*, 1573-1579.

[46] Vilar, S.; Cozza, G.; Moro, S. Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. *Curr. Top. Med. Chem.*, **2008**, *8*, 1555-1572.

[47] Prado-Prado, F.J.; Garcia-Mera, X.; Gonzalez-Diaz, H. Multi-target spectral moment QSAR *vs* ANN for antiparasitic drugs against different parasite species. *Bioorg. Med. Chem.*, **2010**, *18*, 2225-2231.

[48] Gonzalez-Diaz, H. Network topological indices, drug metabolism, and distribution. *Curr. Drug Metab.*, *11*, 283-284.

[49] Khan, M.T. Predictions of the ADMET properties of candidate drug molecules utilizing different QSAR/QSPR modelling approaches. *Curr. Drug Metab.*,**2010**, *11*, 285-295.

[50] Mrabet, Y.; Semmar, N. Mathematical methods to analysis of topology, functional variability and evolution of metabolic systems based on different decomposition concepts. *Curr. Drug Metab.*, **2010**, *11*, 315-341.

[51] Martinez-Romero, M.; Vazquez-Naya, J.M.; Rabunal, J.R.; Pita-Fernandez, S.; Macenlle, R.; Castro-Alvarino, J.; Lopez-Roses, L.; Ulla, J.L.; Martinez-Calvo, A.V.; Vazquez, S.; Pereira, J.; Porto-Pazos, A.B.; Dorado, J.; Pazos, A.; Munteanu, C.R. Artificial intelligence techniques for colorectal cancer drug metabolism: ontology and complex network. *Curr. Drug Metab.*, **2010**, *11*, 347-368.

[52] Zhong, W.Z.; Zhan, J.; Kang, P.; Yamazaki, S. Gender specific drug metabolism of PF-02341066 in rats--role of sulfoconjugation. *Curr. Drug Metab.*, **2010**, *11*, 296-306.

[53] Wang, J.F.; Chou, K.C. Molecular modeling of cytochrome P450 and drug metabolism. *Curr. Drug Metab.*, **2010**, *11*, 342-346.

[54] Gonzalez-Diaz, H.; Duardo-Sanchez, A.; Ubeira, F.M.; Prado-Prado, F.; Perez-Montoto, L.G.; Concu, R.; Podda, G.; Shen, B. Review of MARCH-INSIDE & complex networks prediction of drugs: ADMET, anti-parasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers. *Curr. Drug Metab.*, **2010**, *11*, 379-406.

[55] Garcia, I.; Diop, Y.F.; Gomez, G. QSAR & complex network study of the HMGR inhibitors structural diversity. *Curr. Drug Metab.*, **2010**, *11*, 307-314.

[56] Chou, K.C. Graphic rule for drug metabolism systems. *Curr. Drug Metab.*, **2010**, *11*, 369-378.

[57] Concu, R.; Podda, G.; Ubeira, F.M.; Gonzalez-Diaz, H. Review of QSAR models for enzyme classes of drug targets: theoretical background and applications in parasites, hosts, and other organisms. *Curr. Pharm. Des.*, **2010**, *16*, 2710-2723.

[58] Estrada, E.; Molina, E.; Nodarse, D.; Uriarte, E. Structural contributions of substrates to their binding to P-glycoprotein. A TOPS-MODE approach. *Curr. Pharm. Des.*, **2010**, *16*, 2676-2709.

[59] Garcia, I.; Fall, Y.; Gomez, G. QSAR, Docking, and CoMFA studies of GSK3 inhibitors. *Curr. Pharm. Des.*, **2010**, *16*, 2666-2675.

[60] González-Díaz, H. QSAR and complex networks in pharmaceutical design, microbiology, parasitology, toxicology, cancer, and neurosciences. *Curr. Pharm. Des.*, **2010**, *16*, 2598-2600.

[61] Gonzalez-Díaz, H.; Romaris, F.; Duardo-Sanchez, A.; Perez-Mototo, L.G.; Prado-Prado, F.; Patlewicz, G.; Ubeira, F.M. Predicting drugs and proteins in parasite infections with topological indices of complex networks: theoretical backgrounds, aplications, and legal issues. *Curr. Pharm. Des.*, **2010**, *16*, 2737-2764.

[62] Marrero-Ponce, Y.; Casanola-Martin, G.M.; Khan, M.T.; Torrens, F.; Rescigno, A.; Abad, C. Ligand-based computer-aided discovery of tyrosinase inhibitors. Applications of the TOMOCOMD-CARDD method to the elucidation of new compounds. *Curr. Pharm. Des.*, **2010**, *16*, 2601-2624.

[63] Munteanu, C.R.; Fernandez-Blanco, E.; Seoane, J.A.; Izquierdo-Novo, P.; Rodriguez-Fernandez, J.A.; Prieto-Gonzalez, J.M.; Rabunal, J.R.; Pazos, A. Drug discovery and design for complex diseases through

QSAR computational methods. *Curr. Pharm. Des.*, **2010**, *16*, 2640-2655.

[64] Roy, K.; Ghosh, G. Exploring QSARs with extended topochemical atom (ETA) indices for modeling chemical and drug toxicity. *Curr. Pharm. Des.*, **2010**, *16*, 2625-2639.

[65] Speck-Planche, A.; Scotti, M.T.; de Paulo-Emerenciano, V. Current pharmaceutical design of antituberculosis drugs: future perspectives. *Curr. Pharm. Des.*, **2010**, *16*, 2656-2665.

[66] Vazquez-Naya, J.M.; Martinez-Romero, M.; Porto-Pazos, A.B.; Novoa, F.; Valladares-Ayerbes, M.; Pereira, J.; Munteanu, C.R.; Dorado, J. Ontologies of drug discovery and design for neurology, cardiology and oncology. *Curr. Pharm. Des.*, **2010**, *16*, 2724-2736.

[67] Chakrabarti, E.; Ghosh, S.; Sadhukhan, S.; Sayre, L.; Tochtrop, G. P.; Smith, J.D. Synthesis and biological evaluation of analogues of a novel inhibitor of beta-amyloid secretion. *J. Med. Chem.*, **2010**, *53*, 5302-5319.

[68] Malamas, M.S.; Erdei, J.; Gunawan, I.; Turner, J.; Hu, Y.; Wagner, E.; Fan, K.; Chopra, R.; Olland, A.; Bard, J.; Jacobsen, S.; Magolda, R.L.; Pangalos, M.; Robichaud, A.J. Design and synthesis of 5,5'-disubstituted aminohydantoins as potent and selective human beta-secretase (BACE1) inhibitors. *J. Med. Chem.*, **2010**, *53*, 1146-1158.

[69] Back, M.; Nyhlen, J.; Kvarnstrom, I.; Appelgren, S.; Borkakoti, N.; Jansson, K.; Lindberg, J.; Nystrom, S.; Hallberg, A.; Rosenquist, S.; Samuelsson, B. Design, synthesis and SAR of potent statine-based BACE-1 inhibitors: exploration of P1 phenoxy and benzyloxy residues. *Bioorg. Med. Chem.*, **2008**, *16*, 9471-9486.

[70] Lindsley, S.R.; Moore, K.P.; Rajapakse, H.A.; Selnick, H.G.; Young, M.B.; Zhu, H.; Munshi, S.; Kuo, L.; McGaughey, G.B.; Colussi, D.; Crouthamel, M.C.; Lai, M.T.; Pietrak, B.; Price, E.A.; Sankaranarayanan, S.; Simon, A.J.; Seabrook, G.R.; Hazuda, D.J.; Pudvah, N.T.; Hochman, J.H.; Graham, S.L.; Vacca, J.P.; Nantermet, P.G. Design, synthesis, and SAR of macrocyclic tertiary carbinamine BACE-1 inhibitors. *Bioorg. Med. Chem. Lett.*, **2007**, *17*, 4057-4061.

[71] Stachel, S.J.; Coburn, C.A.; Steele, T.G.; Crouthamel, M.C.; Pietrak, B.L.; Lai, M.T.; Holloway, M.K.; Munshi, S.K.; Graham, S.L.; Vacca, J.P. Conformationally biased P3 amide replacements of beta-secretase inhibitors. *Bioorg. Med. Chem. Lett.*, **2006**, *16*, 641-644.

[72] Garcia, I.; Diop, Y.F.; Gomez, G. QSAR & complex network study of the HMGR inhibitors structural diversity. *Curr. Drug Metab.*, **2010**, *11*, 307-314.

[73] Al-Nadaf, A.; Abu Sheikha, G.; Taha, M.O. Elaborate ligand-based pharmacophore exploration and QSAR analysis guide the synthesis of novel pyridinium-based potent beta-secretase inhibitory leads. *Bioorg. Med. Chem.*, **2010**, *18*, 3088-3115.

[74] Pandey, A.; Mungalpara, J.; Mohan, C.G. Comparative molecular field analysis and comparative molecular similarity indices analysis of hydroxyethylamine derivatives as selective human BACE-1 inhibitor. *Mol. Divers.*, **2010**, *14*, 39-49.

[75] Polgar, T.; Keseru, G.M. Virtual screening for beta-secretase (BACE1) inhibitors reveals the importance of protonation states at Asp32 and Asp228. *J. Med. Chem.*, **2005**, *48*, 3749-3755.

[76] Hetenyi, C.; Paragi, G.; Maran, U.; Timar, Z.; Karelson, M.; Penke, B. Combination of a modified scoring function with two-dimensional descriptors for calculation of binding affinities of bulky, flexible ligands to proteins. *J. Am. Chem. Soc.*, **2006**, *128*, 1233-1239.

[77] Moitessier, N.; Therrien, E.; Hanessian, S. A method for induced-fit docking, scoring, and ranking of flexible ligands. Application to peptidic and pseudopeptidic beta-secretase (BACE 1) inhibitors. *J. Med. Chem.*, **2006**, *49*, 5885-5894.

[78] Salmon, S. A.; Watts, J.L. Minimum inhibitory concentration determinations for various antimicrobial agents against 1570 bacterial isolates from turkey poults. *Avian Diseases*, **2000**, *44*, 85-98.

[79] Chou, K.C. Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.*, **2004**, *11*, 2105-2134.

[80] Chou, K.C.; Wei, D.Q.; Du, Q.S.; Sirois, S.; Zhong, W.Z. Progress in computational approach to drug development against SARS. *Curr. Med. Chem.*, **2006**, *13*, 3263-3270.

[81] Prado-Prado, F.J.; Gonzalez-Diaz, H.; de la Vega, O.M.; Ubeira, F.M.; Chou, K.C. Unified QSAR approach to antimicrobials. Part 3: first multi-tasking QSAR model for input-coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. *Bioorg. Med. Chem.*, **2008**, *16*, 5871-5880.

[82] Prado-Prado, F.J.; de la Vega, O.M.; Uriarte, E.; Ubeira, F.M.; Chou, K.C.; Gonzalez-Diaz, H. Unified QSAR approach to antimicrobials. 4. Multi-target QSAR modeling and comparative multi-distance study of the giant components of antiviral drug-drug complex networks. *Bioorg. Med. Chem.*, **2009**, *17*, 569-575.

[83] Kubinyi, H. Quantitative structure-activity relationships (QSAR) and molecular modelling in cancer research. *J. Cancer Res. Clin. Oncol.*, **1990**, *116*, 529-537.

[84] Prado-Prado, F.J.; Borges, F.; Perez-Montoto, L.G.; Gonzalez-Diaz, H. Multi-target spectral moment: QSAR for antifungal drugs *vs* different fungi species. *Eur. J. Med. Chem.*, **2009**, *44*, 4051-4056.

[85] Estrada, E. On the topological sub-structural molecular design (TOSS-MODE) in QSPR/QSAR and drug design research. *SAR QSAR Environ. Res.*, **2000**, *11*, 55-73.

[86] Overington, J. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. *J. Comput. Aided Mol. Des.*, **2009**, *23*, 195-198.

[87] Estrada, E.; Patlewicz, G.; Chamberlain, M.; Basketter, D.; Larbey, S. Computer-aided knowledge generation for understanding skin sensitization mechanisms: the TOPS-MODE approach. *Chem. Res. Toxicol.*, **2003**, *16*, 1226-1235.

[88] Estrada, E.; Uriarte, E.; Gutierrez, Y.; Gonzalez, H. Quantitative structure-toxicity relationships using TOPS-MODE. 3. Structural factors influencing the permeability of commercial solvents through living human skin. *SAR QSAR Environ. Res.*, **2003**, *14*, 145-163.

[89] Estrada, E.; Gonzalez, H. What are the limits of applicability for graph theoretic descriptors in QSPR/QSAR? Modeling dipole moments of aromatic compounds with TOPS-MODE descriptors. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 75-84.

[90] Estrada, E.; Vilar, S.; Uriarte, E.; Gutierrez, Y. *In silico* studies toward the discovery of new anti-HIV nucleoside compounds with the use of TOPS-MODE and 2D/3D connectivity indices. 1. Pyrimidyl derivatives. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 1194-1203.

[91] Estrada, E. Quantum-chemical foundations of the topological substructural molecular design. *J. Phys. Chem. A*, **2008**, *112*, 5208-5217.

[92] Yoshii, F.; Hirono, S. Construction of a quantitative three-dimensional model for odor quality using comparative molecular field analysis (CoMFA). *Chem. Senses*, **1996**, *21*, 201-210.

[93] Gonzalez-Diaz, H.; Saiz-Urra, L.; Molina, R.; Santana, L.; Uriarte, E. A model for the recognition of protein kinases based on the entropy of 3D van der Waals interactions. *J. Proteome Res.*, **2007**, *6*, 904-908.

[94] Alvarez-Ginarte, Y.M.; Marrero-Ponce, Y.; Ruiz-Garcia, J.A.; Montero-Cabrera, L.A.; Vega, J.M.; Noheda Marin, P.; Crespo-Otero, R.; Zaragoza, F.T.; Garcia-Domenech, R. Applying pattern recognition methods plus quantum and physico-chemical molecular descriptors to analyze the anabolic activity of structurally diverse steroids. *J. Comput. Chem.*, **2008**, *29*, 317-333.

[95] Morales, A.H.; Rodriguez-Borges, J.E.; García-Mera, X.; Fernández, F.; Dias-Sueiro-Cordeiro, M.N. Probing the anticancer activity of nucleoside analogues: a QSAR model approach using an internally consistent training set. *J. Med. Chem.*, **2007**, *50*, 1537-1545.

Received: March 30, 2011 **Revised: June 5, 2011** **Accepted: September 12, 2011**

Current Bioinformatics **2011**, 0, 000-000

Review of Bioinformatics and Theoretical studies of Acetylcholinesterase inhibitors

Manuel Escobar, Franco Fernández, Xerardo García-Mera and Francisco Prado-Prado*

Department of Organic Chemistry, University of Santiago de Compostela, 15782, Spain.

Abstract: Alzheimer's disease is a complex disease, and no single "magic bullet" is likely to prevent or cure it. That's why current treatments focus on several different aspects, including helping people maintain mental function; managing behavioral symptoms; and slowing, delaying, or preventing the disease. Four medications are approved by the U.S. Food and Drug Administration to treat Alzheimer's. Donepezil, rivastigmine, and galantamine are used to treat mild to moderate Alzheimer's. Memantine is used to treat moderate to severe Alzheimer's. These drugs work by regulating neurotransmitters (the chemicals that transmit messages between neurons). Treatment of AD by ACh precursors and cholinergic agonists was ineffective or caused severe side effects. ACh hydrolysis by AChE causes termination of cholinergic neurotransmission. Therefore, compounds which inhibit AChE might significantly increase the levels of ACh depleted in AD. However, these drugs don't change the underlying disease process and may help only for a few months to a few years. In this sense, quantitative structure-activity relationships (QSAR) could play an important role in studying these AChE inhibitors. QSAR models are necessary in order to guide the AChE synthesis. In this work, we revised different bioinformatics and theoretical studies of Acetylcholinesterase inhibitors, design and computational studies for a very large and heterogeneous series of AChE inhibitors. First, we review 2D QSAR, 3D QSAR, CoMFA, CoMSIA and Docking and new theoretical methodology with different compound to find out the structural requirements. Next, we revised QSAR studies using method of Linear Discriminant Analysis (LDA) in order to understand the essential structural requirement for binding with receptor for AChE inhibitors.

Keywords: QSAR; CoMSIA; COMFA; topological indices; Molecular Docking; Acetylcholinesterase inhibitors; Alzheimer's disease (AD).

INTRODUCTION

Alzheimer's disease (AD) is a disorder that attacks the central nervous system through progressive degeneration of its neurons. AD occurs in around 10% of the elderly and, as yet, there is no known cure. Patients with this disease develop dementia which becomes more severe as the disease progresses [1]. It was suggested that symptoms of AD are caused by decrased of activity of cholinergicneocortical and hippocampal neurons. Treatment of AD by ACh precursors and cholinergic agonists was ineffective or caused severe side effects. ACh hydrolysis by AChE causes termination of cholinergic neurotransmission. Therefore, compounds which inhibit AChE might significantly increase the levels of ACh depleted in AD. Indeed, it was shown that AChE inhibitors improve the cognitive abilities of AD patients at early stages of the disease development [2-4].

Acetylcholinesterase (AChE) is key enzyme in the nervous system of animals. By rapid hydrolysis of the neurotransmitter, acetylcholine (ACh), AChE terminates neurotransmission at cholinergic synapses. It is a very fast enzyme, especially for a serine hydrolase, functioning at a rate approaching that of a diffusion-controlled reaction.

***Corresponding author:** Prado-Prado, Francisco, Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, 15782, Email: francisco.prado@usc.es.

AChE inhibitors are among the key drugs approved by the FDA for management of Alzheimer's disease (AD) [5-8]. The powerful toxicity of organophosphorus (OP) poisons is attributed primarily to their potent AChE inhibitors.

Solution of the three-dimensional (3D) structure of Torpedo californica acetylcholinesterase (TcAChE) in 1991 opened up new horizons in research on an enzyme that had already been the subject of intensive investigation [9]. The unanticipated structure of this extremely rapid enzyme, in which theactive site was found to be buried at the bottom of a deep and narrow gorge, lined by 14 aromatic residues (colored dark magenta), led to a revision of the views then held concerning substrate traffic, recognition and hydrolysis [10]. To understand how those aromatic residues behave with the enzyme, see Flexibility of aromatic residues in acetylcholinesterase .

Solution of the 3D structure of acetylcholinesterase led to a series of theoretical and experimental studies, which took advantage of recent advances in theoretical techniques for treatment of proteins or enzimes, such as molecular dynamics and electrostatics and to site-directed mutagenesis, utilizing suitable expression systems. Acetylcholinesterase hydrolysizes theneurotra nsmitter acetylcholine (ACh), producing choline and an acetate group. Acetylcholine directly binds Ser200(via its nucleophilic Oγ atom) within the catalytic triad (Ser200, His440, and Glu327) (ACh/TcAChE structure 2ace). The residues Trp84 and Phe330 are also

1574-8936/11 $55.00+.00 ©

important in the ligand recognition. See also: AChE inhibitors and substrates, see **Figure 1**.

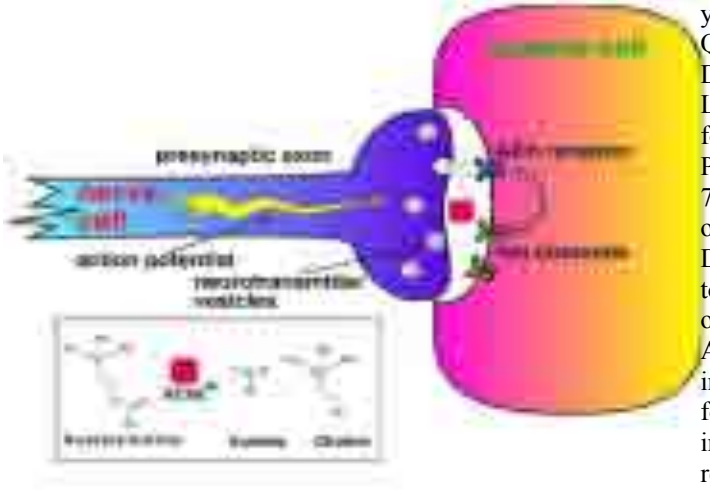


Figure 1. Cholinergic Synapse: Key Enzyme in the Nervous System.

In this part, Chemoinformatics and Bioinformatics methods may play an important role in the study of acetylcholinesterase inhibitors, Quantitative Structure-Activity Relationships (QSAR) studies are used as predictive tools for the molecular development [11, 12]. Up to today, there are near 1600 molecular descriptors that, in principle, can be generalized and used to solve the former problem [13]. Many of these indices are known as molecular Topological Indices (TIs) or simply invariants of a molecular graph. Unfortunately, QSAR studies are generally based on databases considering only structurally parent compounds acting against one single microbial species. In a recent review, our group have discussed recent advances in the field [14]. In addition to QSAR, Bioinformatics and Chemoinformatics methods useful to study β-secretase may include techniques like Comparative Molecular Field Analysis (CoMFA), drug-target Docking, Sequence Alignment (SA) or other methods. In a recent, preliminary review in the field published in Proteomics in 2008 was discussed the use of these methods but only from the point of view of proteins [15]. Almost all QSAR techniques are based on the use of molecular descriptors, which are numerical series that codify useful chemical information and enable correlations between statistical and biological properties [16, 17]. On the other hand, QSAR models can be used to explore the relationships between the structural spaces of compounds as inhibitors for specific enzymes, such as MAO inhibitors [18], HIV-1 integrase inhibitors [19], and/or protease inhibitors [20] or tyrosinase inhibitors [21-23]. In fact, recently, the field has moved from small molecules to proteins and other systems. For instance, González-Díaz *et al.* discussed the use of these methods but only from the point of view of proteins [15]. Later, some groups published different papers in one special issue on QSAR but also restricted to the field of protein and proteomics [24-30]. In other recent issue, guest-edited by González-Díaz [31] appeared a series of papers devoted to QSAR/QSPR techniques for low-molecular-weight drugs [31-40]. Most recently, Prado-Prado *et al.* [41] published a mt-QSAR for anti-parasitic drugs. This year was published other issue [42] focused on QSAR/QSPR models and graph theory used to approach Drug ADMET processes and Metabolomics [43-50]. Last, two of the most recent issues published have been focused on the discussions of the applications of QSAR in Pharmaceutical Design [51-60] and Bioinformatics [61-70]. In the present work, we review previous works based on 2D-QSAR, 3D-QSAR, CoMFA, CoMSIA and Docking techniques, which studied different compounds to find out the structural requirements. Last, we carried out new QSAR studies using Linear Discriminant Analysis (LDA) method and the software ModesLab[71] in order to understand the essential structural requirement for binding with receptor for *acetylcholinesterase* inhibitors. The topics reviewed, discussed, and/or reported in this paper are:

1. Theoretical studies for acetylcholinesterase inhibitors.

1.1. Acaricidal and QSAR of monoterpenes against Tetranychus urticae.
1.2. Preparation, AChE activity, docking study, and 3D-QSAR.
1.3. Prediction of AChE inhibitors and characterization by machine learning methods.
1.4. 3D-QSAR Studies of Physostigmine Analogues as AChE Inhibitors.
1.5. 3D-QSAR Studies on Carbamates as ACeH Inhibitors.
1.6. Classification of drug considering their IC50 using linear programming.
1.7. An investigation of carbamates for AChE inhibition using 3D-QSAR.
1.8. Molecular docking and 3D-QSAR studies indanone as AChE inhibitors.
1.9. QSAR of tacrine derivatives against AChE activity using variable selections.
1.10. 3D QSAR studies of AChE inhibitors based on molecular docking.
1.11. Anchor-GRIND: Filling the Gap between Standard 3D QSAR-GRIND.
1.12. A Docking Score Function for Estimating Ligand-Protein Interactions.
1.13. Modulation of Binding Strength in Several Classes of Active Site Inhibitors.
1.14. Structure-based 3D QSAR and design of novel ACeH inhibitors.

2. New method for the study of new acetylcholinesterase inhibitors.

2.1. Preface to new QSAR study acetylcholinesterase inhibitors
2.2. Methods
2.3. Results and Discussion.

1. Theoretical studies for acetylcholinesterase inhibitors.

In this section we updated the contents presented in our recent review published in Current Drugs

Metabolim [72]. The high number of possible candidates to acetylcholinesterase inhibitors creates the necessity of Quantitative Structure-Activity Relationship models in order to guide the acetylcholinesterase inhibitor synthesis. In this work, we revised different computational studies for a very large and heterogeneous series of acetylcholinesterase. First, we review 3D QSAR, CoMFA, CoMSIA and Docking with different compound to find out the structural requirements for acetylcholinesterase inhibitors. Last, we carried out new QSAR studies using Linear Discriminant Analysis (LDA) method and the software ModesLab[71] in order to understand the essential structural requirement for binding with receptor for *acetylcholinesterase* inhibitors.

1.1. Acaricidal and QSAR of monoterpenes against Tetranychus urticae.

Badawy *et al.* [73] reviewed the acaricidal activity of 12 monoterpenes against the two-spotted spider mite, Tetranychus urticae Koch, was examined using fumigation and direct contact application methods. Cuminaldehyde and (-)-linalool showed the highest fumigant toxicity with LC_{50} = 0.31 and 0.56 mg/l, respectively. The other monoterpenes exhibited a strong fumigant toxicity, the LC_{50} values ranging from 1.28 to 8.09 mg/l, except camphene, which was the least effective (LC_{50} = 61.45 mg/l). Based on contact activity, the results were rather different: menthol displayed the highest acaricidal activity (LC50 = 128.53 mg/l) followed by thymol (172.0 mg/l), geraniol (219.69 mg/l) and (-)-limonene (255.44 mg/l); 1-8-cineole, cuminaldehyde and (-)-linalool showed moderate toxicity. At 125 mg/l, (-)-Limonene and (-)-carvone caused the highest egg mortality among the tested compounds (70.6 and 66.9% mortality, respectively). In addition, the effect of molecular descriptors was also analyzed using the quantitative structure activity relationship (QSAR) procedure. The QSAR model showed excellent agreement between the estimated and experimentally measured toxicity parameter (LC_{50}) for the tested monoterpenes and the fumigant activity increased significantly with the vapor pressure. The authors compared the results of the fumigant and contact toxicity assays of monoterpenes against T. urticae with the results of acetylcholinesterase (AChE) inhibitory effect revealed that some of the tested compounds showed a strong acaricidal activity and a potent AChE inhibitory activity, such as cuminaldehyde, (-)-linalool, (-)-limonene and menthol. However, other compounds such as (-)-carvone revealed a strong fumigant activity but a weak AChE inhibitory activity.

1.2. Preparation, AChE activity, docking study, and 3D-QSAR

Synthesis and anticholinesterase activity of 4-aryl-4-oxo-N-phenyl-2-aminylbutyramides, novel class of reversible, moderately potent cholinesterase inhibitors, are reported by Vitorovic-Todorovic *et al.*[74]. In this work the authors used a simple substituent variation on aroyl moiety changes anti-AChE activity for two orders of magnitude; also substitution and type of hetero (ali) cycle in position 2 of butanoic moiety govern AChE/BChE selectivity. The most potent compounds showed mixed-type inhibition, indicating their binding to free enzyme and enzyme–substrate complex.

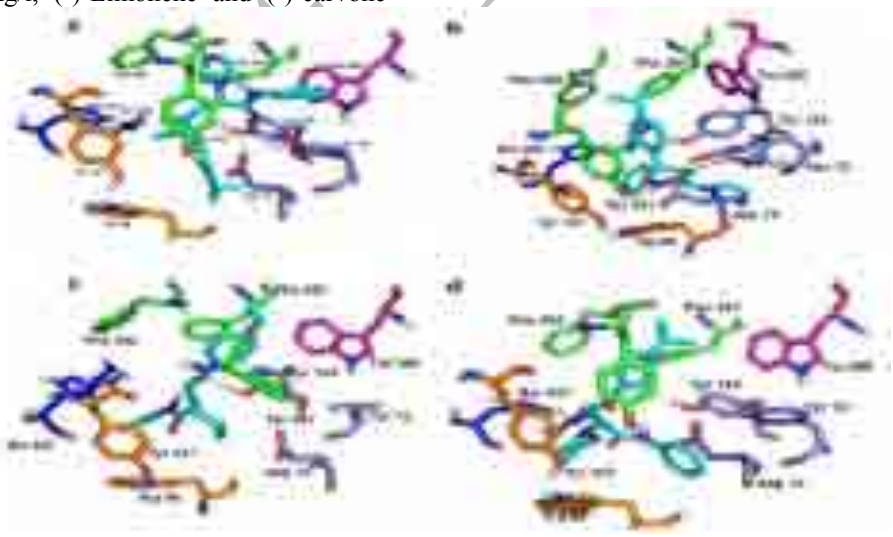


Figure 2. Compounds 5(R), (a); 5(S), (b); 19(R), (c); and 19(S), (d), docked into the binding site of the AChE, highlightening the protein residues that form the main interactions with the different structural units of the inhibitors. Hydrogen bonds are represented as black dots. Residues are colored as follows: anionic site Trp 86 and Tyr 337, orange; PAS Tyr 72, Asp 74, Tyr 124, Ser 125 violet and Trp 286 magenta; acyl binding pocket Tyr 341, Phe 297 and Phe 338 green; catalytic triad residues His 447 and Ser 203 dark blue; ligand, turquoise.

Alignment-independent 3D QSAR study on reported compounds, and compounds having similar potencies obtained from the literature, confirmed that alkyl substitution on aroyl moiety of molecules is requisite for inhibition activity. The presence of hydrophobic moiety at close distance from hydrogen bond acceptor has favorable influence on inhibition potency. Docking studies show that compounds probably bind in the middle of the AChE active site gorge, but are buried deeper inside BChE active site gorge, as a consequence of larger BChE gorge void, see **Figure 2**.

1.3. Prediction of AChE inhibitors and characterization by machine learning methods.

Acetylcholinesterase (AChE) has become an important drug target and its inhibitors have proved useful in the symptomatic treatment of Alzheimer's disease. This work Wei Lv *et al.* [75] explores several machine learning methods (support vector machine (SVM), k-nearest neighbor (k-NN), and C4.5 decision tree (C4.5 DT)) for predicting AChE inhibitors (AChEIs). A feature selection method is used for improving prediction accuracy and selecting molecular descriptors responsible for distinguishing AChEIs and non-AChEIs. The prediction accuracies are 76.3%w88.0% for AChEIs and 74.3%w79.6% for non-AChEIs based on the three kinds of machine learning methods. This work suggests that machine learning methods such as SVM are facilitating for predicting AChEIs potential of unknown sets of compounds and for exhibiting the molecular descriptors associated with AChEIs, see **Figure 3**.

Figure 3. The structures of the misclassified AChEIs.

1.4. 3D-QSAR Studies of Physostigmine Analogues as AChE Inhibitors.

Natural alkaloid Physostigmine is one of the most potent pseudo-irreversible inhibitor of Acetylcholinesterase. It was found to accelerate long-term memory process, but due to its short half life and variable bioavailability, has inconsistent clinical efficacy. Zaheer Ul-Haq *et al.* [76] proposed a 3D-QSAR studies based on the comparative molecular field analysis and comparative molecular similarity indices analysis and were applied to a set of 40 Physostigmine derivatives which are divided into two classes: A and B. In this study was obtained a highly reliable and extensive dynamic QSAR model based on alignment procedure with co-crystallized Ganstigmine as template. The strategy yielded significant 3DQSAR models with the cross-validated q^2 values 0.762 and 0.754 for comparative molecular field analysis and comparative molecular similarity indices analysis, respectively. Resulted models were validated by external set of eight compounds yielding high correlation coefficient r^2 values of 0.730 and 0.720 for comparative molecular field analysis and comparative molecular similarity indices analysis, respectively. Furthermore, the analysis of comparative molecular field analysis and comparative molecular similarity indices analysis contour maps within the active site of AChE were conducted in order to understand the interactions between the receptor and the Physostigmine derivatives, see **Figure 4**.

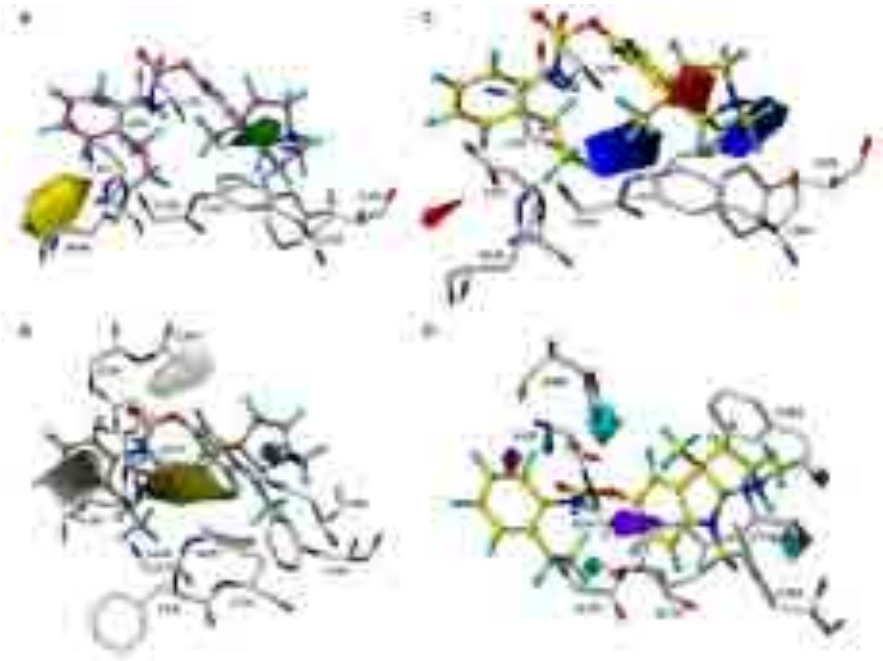


Figure 4. 3D CoMSIA steric (A), CoMSIA electrostatic (B), CoMSIA hydrophobic (C), CoMSIA H-bond donor and acceptor (D) contour maps.

This class of study will facilitate the rational design of more potent Physostigmine compounds which might have better activity and reduce toxicity for the treatment of Alzheimer disease.

1.5. 3D-QSAR Studies on Carbamates as ACeH Inhibitors

In view of the nonavailability of complete X-ray structure of carbamates cocrystallized with AChE enzyme, the 3D-QSAR model development based on cocrystallized conformer (CCBA) as well as docked conformerbased alignment (DCBA) is not feasible. Therefore, the only two alternatives viz. pharmacophore and maximum common substructure-based alignments are left for the 3D-QSAR comparative molecular field analyses (CoMFA) and comparative molecular similarity indices analyses (CoMSIA) model development. Shailendra S. Chaudhaery *et al.* [77] in the presents iin this study, a 3D-QSAR models that have been developed using both alignment methods, where CoMFA and CoMSIA models based on pharmacophore-based alignment were in good agreement with each other and demonstrated significant superiority over MCS-based alignment in terms of leave-one-out (LOO) cross-validated q^2 values of 0.573 and 0.723 and the r^2 values of 0.972 and 0.950, respectively. The validation of the best CoMFA and CoMSIA models based on pharmacophore (Hip-Hop)-based alignment on a test set of 17 compounds provided significant predictive r^2 [r^2_{pred}(test)] of 0.614 and 0.788, respectively. The authors presented a contour map analyses, that revealed the relative importance of steric, electrostatic, and hydrophobicity for AChE inhibition activity, see **Figure 5**. However, hydrophobic factor plays a major contribution to the AChE inhibitory activity modulation which is in strong agreement with the fact that the AChE is having a wide active site gorge (~20 Å) occupied by a large number of hydrophobic amino acid residues.

Figure 5. The docked bioactive conformation of the most active compound 1 into the active site of the AChE enzyme generated by GOLD (version 3.0.1). Representation of hydrophobic (A) and electrostatic (B) zones around the most active compound 1.

1.6. Classification of drug considering their IC_{50} using linear programming

A priori analysis of the activity of drugs on the target protein by computational approaches can be useful in narrowing down drug candidates for further experimental tests. Currently, there are a large number of computational methods that predict the activity of drugs on proteins. In this study, Pelin Armutlu *et al.* [78] approach the activity prediction problem as a classification problem and, they aim to improve the classification accuracy by introducing an algorithm that combines partial least squares regression with mixed-integer programming based hyper-boxes classification method, where drug molecules are classified as low active or high active regarding their binding activity (IC_{50} values) on target proteins. In this work they also aim to determine the most significant molecular

descriptors for the drug molecules, see **Figure 6**. First analyzing the activities of widely known inhibitor datasets including Acetylcholinesterase (ACHE), Benzodiazepine Receptor (BZR), Dihydrofolate Reductase (DHFR), Cyclooxygenase-2 (COX-2) with known IC_{50} values. The results at this stage proved that their approach consistently gives better classification accuracies compared to 63 other reported classification methods such as SVM, Naïve Bayes, where they were able to predict the experimentally determined IC_{50} values with a worst case accuracy of 96%. To further test applicability of this approach first the authors created dataset for Cytochrome P450 C17 inhibitors and then predicted their activities with 100% accuracy. The authors concluded results indicate that this approach can be utilized to predict the inhibitory effects of inhibitors based on their molecular descriptors.

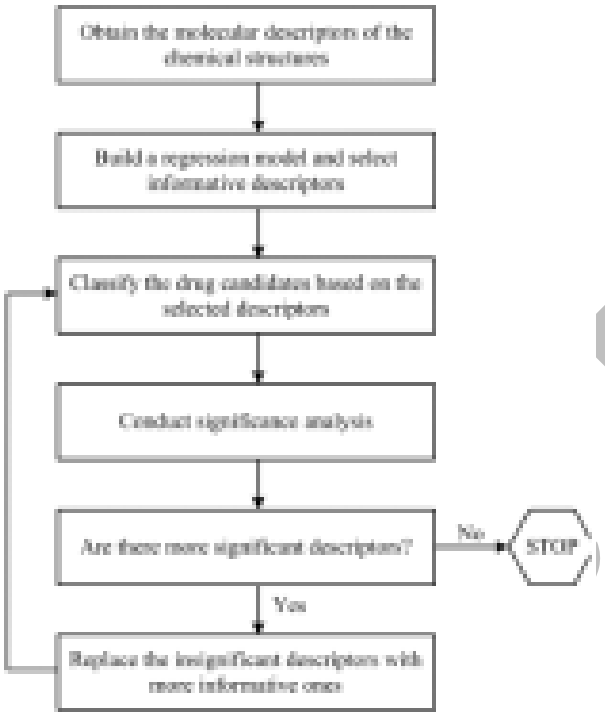


Figure 6. Outline of classification approach.

1.7. An investigation of carbamates for AChE inhibition using 3D-QSAR analysis

Kuldeep K. Roy *et al.* in order to identify the essential structural features and physicochemical properties for AChE inhibitory activity in some carbamate derivatives, the systematic QSAR studies (CoMFA, advance CoMFA and CoMSIA) have been carried out on a series of (total 78 molecules) taking 52 and 26 molecules in training and test set, respectively. Statistically significant 3D-QSAR (three-dimensional Quantitative Structure Activity Relationship) models were developed on training setmolecules using CoMFA and CoMSIA and validated against test set compounds. The highly predictive models (CoMFA $q^2 = 0.733$, $r^2 = 0.967$, predictive $r^2 = 0.732$, CoMSIA $q^2 = 0.641$, $r^2 = 0.936$, predictive $r^2 = 0.812$) well explained the variance in binding affinities both for the training and the test set compounds. The generated models suggest that steric, electrostatic and hydrophobic interactions play an important role in describing the variation in binding affinity. In particular the carbamoyl nitrogen should be more electropositive; substitutions on this nitrogen should have high steric bulk and hydrophobicity while the amino nitrogen should be electronegative in order to have better activity, see **Figure 7**. These studies may provide important insights into structural variations leading to the development of novel AChE inhibitorswhich may be useful in the development of novelmolecules for the treatment of Alzheimer's disease.

Figure 7. Contour map of the best CoMFA model using Tripos standard field and contour map of the best CoMSIA model using steric, electrostatic and hydrophobic fields.

1.8. Molecular docking and 3D-QSAR studies indanone as AChE inhibitors

Liang-liang Shen *et al.*[79] explore the binding mode of 2-substituted 1-indanone derivatives with acetylcholinesterase (AChE) and provide hints for the future design of new derivatives with higher potency and specificity. The GOLD-docking conformations of the compounds in the active site of the enzyme were used in subsequent studies were the method used by the autors. The highly reliable and predictive 3D-QSAR models were achieved by comparative molecular field analysis (CoMFA) and comparative molecular similarity analysis (CoMSIA) methods. The predictive capabilities of the models were validated by an external test set. Moreover, the stabilities of the 3D-QSAR models were verified by the leave-4-out cross-validation method. The CoMFA and CoMSIA models were constructed successfully with a good cross-validated coefficient (q^2) and a non-cross-validated coefficient (r^2). The q^2 and r^2 obtained from the leave-1-out cross validation method were 0.784 and 0.974 in the

CoMFA model and 0.736 and 0.947 in the CoMSIA model, respectively. The coefficient isocontour maps obtained from these models were compatible with the geometrical and physicochemical properties of AChE see **Figure 8**.

A conclusion of this method was that the contour map demonstrated the binding affinity could be enhanced when the small protonated nitrogen moiety was replaced by a more hydrophobic and bulky group with a highly partial positive charge. The authors in this study provide a better understanding of the interaction between the inhibitors and AChE, which is helpful for the discovery of new compounds with more potency and selective activity.

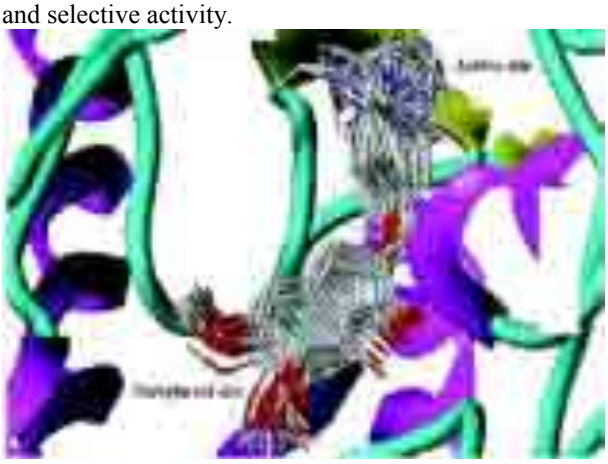


Figure 8. Alignment of all 45 compounds based on GOLD results.

This image was generated with the Base program in SYBYL version 7.0.

1.9. QSAR of tacrine derivatives against AChE activity using variable selections

Jung *et al.*[80] developed a diverse approach to the quantitative structure–activity relationship (QSAR) of tacrine derivatives against acetylcholinesterase (AChE) activity and was studied using variable selections of stepwise multiple linear regression (MLR), genetic algorithm (GA)- MLR, and simulated annealing (SA)-MLR. AChE activity (logRA) of tacrine derivatives was expressed with acceptable explanation (95.5–95.9%) and good predictive power (94.5–95.2%), respectively, in the models, see **Table 1**.

The best equation was obtained from simulated annealing (SA) MLR with greater explanatory capability and better prediction, with a smaller standard error than other methods. The resulting models with the given descriptors illustrate the significant roles of hydrophobic and electrostatic interaction on increasing AChE activity, but hydrophilic and topological feature of molecules were shown to decrease AChE activity.

Table 1. AChE activity of tacrine derivatives.

Mol. ID	Obs.	Pred.	Pred.	Pred.	Pred.	Resid.	Resid.
1	-1.37	-1.10	-0.99	-1.24	-0.17	0.14	-0.09
2	-1.41	-1.12	-1.48	-0.78	-0.27	-0.37	-0.58
3	-1.41	-1.40	-1.45	-1.24	-0.29	0.06	-0.17
4	-1.23	-1.40	-1.45	-1.19	-0.01	0.04	-0.21
5	-1.19	-1.26	-1.49	-1.19	0.17	0.22	-0.03
6	-1.26	-1.27	-1.44	-1.10	0.07	0.29	-0.08
7	-1.33	-1.15	-1.48	-1.10	0.01	0.17	-0.15
8	-0.71	-1.33	-1.37	-1.24	-0.17	0.14	-0.09
9b	-0.43	-1.36	-1.37	-0.99	0.62	0.65	0.28
10b	-0.68	-1.18	-1.15	-1.16	0.93	0.94	0.72
11b	-0.80	-1.39	-1.51	-0.95	0.50	0.47	0.27
12	-0.85	-1.18	-1.08	-1.19	0.58	0.70	0.39
13	-0.74	-0.65	-0.58	-1.31	0.33	0.23	0.46
14b	-0.74	-1.31	-1.19	-0.88	-0.09	-0.16	0.13

1.10. 3D QSAR studies of AChE inhibitors based on molecular docking and CoMFA

Akula *et al.* [81] were performed a Three-dimensional quantitative structure–activity relationship (3D QSAR) studies on acetylcholinesterase (AChE) inhibitors, based on molecular docking scores obtained by using FlexX and FlexiDock and comparative molecular field analysis (CoMFA). The docking scores were used as molecular descriptors along with the steric and electrostatic field values of CoMFA, for partial least square (PLS) analysis. The high leave one out (LOO) cross-validated correlation coefficient (q^2 = 0.714) reveals that the model is a useful tool for the prediction of test set as well as newly designed structures against AChE activity. The superimposed CoMFA models on the receptor site of AChE are guiding the design of potential inhibitory structures directed against AChE activity, see **Figure 9**.

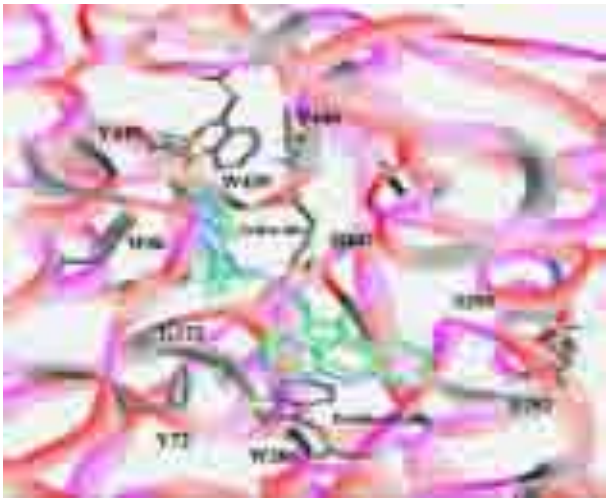


Figure 9. Comparison of the FlexX and FlexiDock, ligand–protein complexes of compound 3B. The bound inhibitor is shown as ball and stick model. The backbone of the protein structure is rendered as shaded ribbon with color by property and the labeled protein residues are in capped stick model with color by atom.

1.11. Anchor-GRIND: Filling the Gap between Standard 3D QSAR and the GRIND

In the present study, Fabien Fontaine *et al.* [82] have introduced the anchor- GRIND (GRid-INdependent Descriptors) method, see **Figure 10**. This method represents a particular application of the GRIND for situations where there is at least a single point in the structure of the considered compounds that can be recognized as common for either chemical or biological reasons. The descriptors obtained are far more specific and therefore produce better models, which are easier to interpret. In addition, the interactions within atoms of the common scaffold are filtered out so that the information contained in the descriptors is highly specific for the sites of variability. The improvement obtained with this new methodology has been demonstrated with three different examples, in which the novel approach allowed to identify in a rather straightforward way the most relevant structural features of active and inactive compounds. The results for the factor Xa data set are particularly interesting, since they show that, when a relevant anchor point can be defined, excellent models can be obtained even for highly diverse compounds. It is also impressive to see how the anchor-GRIND distances match structural features of the factor Xa binding site. Two models were tested, one with two blocks of variables, i.e., the anchor- MIF and the MIF-MIF block, and one with only one block of variables, i.e., the anchor-MIF block. The results for the test set are slightly better for the one block model: 88% of the compounds are well classified versus 84% of the compounds being well classified in the two-block model.

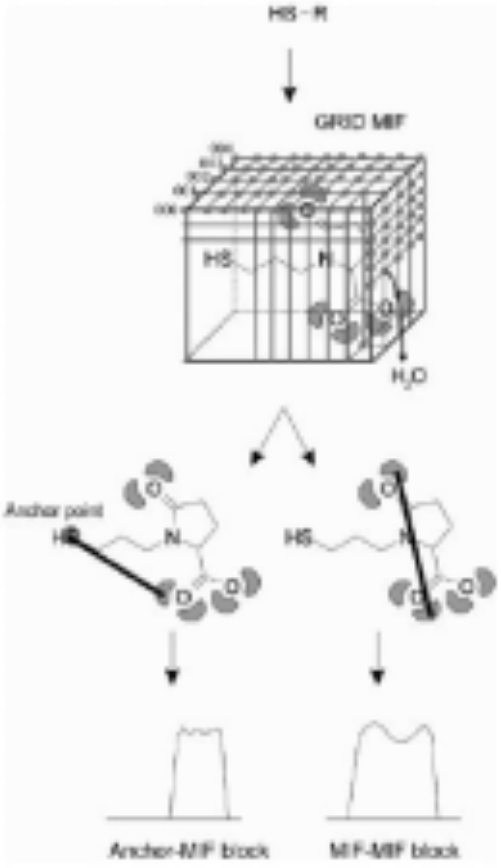


Figure 10. Calculation of the anchor-GRIND descriptors for an ACE inhibitor with the anchor point set on the zinc binder sulfur atom.

1.12. A Docking Score Function for Estimating Ligand-Protein Interactions.

Acetylcholinesterase (AChE) inhibition is an important research topic because of its wide range of associated health implications. A receptor-specific scoring function was developed by Jianxin Guo *et al.* [83] for predicting binding affinities for human AChE (huAChE) inhibitors. This method entails a statistically trained weighted sum of electrostatic and van der Waals (VDW) interactions between ligands and the receptor residues. Within the 53 ligand training set, a strong correlation was found ($r^2 = 0.89$) between computed and experimental inhibition constants. Leave-oneout cross-validation indicated high predictive power ($q^2 = 0.72$), and analysis of a separate 16-compound test set also produced very good correlation with experiment ($r^2 = 0.69$), see **Figure 11**. The authors developed a scoring function analysis that has permitted identification and characterization of important ligand-receptor interactions, producing a list of those residues making the most important electrostatic and VDW contributions within the main active site, gorge area, acyl binding pocket, and peripheral site. These analyses are consistent with X-ray crystallographic and site-directed mutagenesis studies.

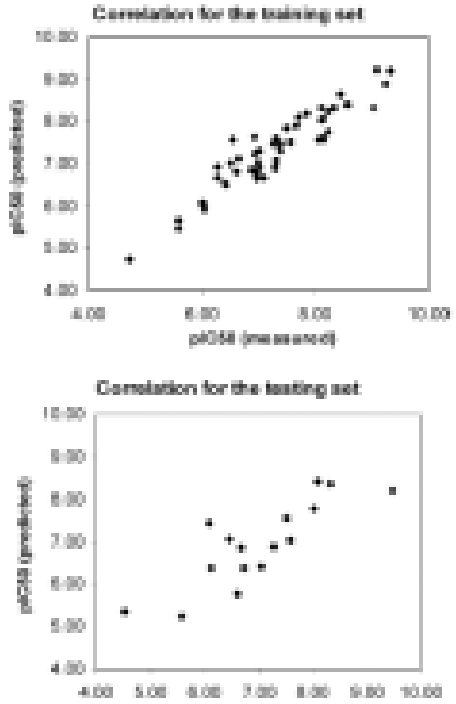


Figure 11. Correlation of the calculated activity (pIC_{50}) with experiment: (top) training set ($r^2 = 0.89$); (bottom) testing set ($r^2 = 0.69$).

1.13. Modulation of Binding Strength in Several Classes of Active Site Inhibitors.

The comparative binding energy (COMBINE) methodology has been used by Martín-Santamaría *et al.* [84] to identify the key residues that modulate the inhibitory potencies of three structurally different classes of acetylcholinesterase inhibitors (tacrines, huprines, and dihydroquinazolines) targeting the catalytic active site of this enzyme. The extended set of energy descriptors and the partial least-squares methodology used by COMBINE analysis on a unique training set containing all the compounds yielded an interpretable model that was able to fit and predict the activities of the whole series of inhibitors reasonably well, $r^2 = 0.91$ and $q^2 = 0.76$, 4 principal components. A more robust model $q^2 = 0.81$ and SDEP = 0.25, 3 principal components) was obtained when the same chemometric analysis was applied to the huprines set alone, but the method was unable to provide predictive models for the other two families when they were treated separately from the rest, the resulst were in **Table 2**. This finding appears to indicate that the enrichment in chemical information brought about by the inclusion of different classes of compounds into a single training set can be beneficial when an internally consistent set of pharmacological data can be derived.

Table 2. Performance of Different COMBINE Models for the Whole Set of Inhibitors

data set*b*	no. of PCs	r^2	SDEC	q^2	SDEP
A	1	0.57	1.03	0.50	1.11
	2	0.70	0.86	0.59	1.00
	3	0.73	0.81	0.59	1.00
	4	0.76	0.76	0.60	0.99
	5	0.83	0.64	0.59	1.00
B	1	0.70	0.86	0.58	1.01
	2	0.79	0.71	0.59	1.00
	3	0.82	0.66	0.61	0.98
	4	0.86	0.58	0.64	0.93
	5	0.89	0.51	0.62	0.97
C	1	0.71	0.85	0.56	1.04
	2	0.79	0.72	0.55	1.05
	3	0.82	0.66	0.55	1.05
	4	0.84	0.62	0.59	1.00
	5	0.86	0.58	0.61	0.97
D	1	0.62	0.97	0.48	1.12
	2	0.72	0.82	0.55	1.05
	3	0.85	0.60	0.64	0.93
	4	**0.91**	**0.47**	**0.76**	**0.78**
	5	0.92	0.44	0.75	0.79
E	1	0.70	0.85	0.57	1.03
	2	0.81	0.68	0.61	0.98
	3	0.84	0.62	0.64	0.94
	4	0.88	0.54	0.68	0.89
	5	0.90	0.50	0.65	0.92
F	1	0.59	1.00	0.51	1.09
	2	0.73	0.82	0.63	0.95
	3	0.76	0.76	0.62	0.97
	4	0.82	0.67	0.62	0.96
	5	0.85	0.61	0.63	0.95
G	1	0.72	0.83	0.61	0.98
	2	0.77	0.74	0.60	0.99
	3	0.82	0.66	0.57	1.02
	4	0.86	0.58	0.56	1.04
	5	0.88	0.53	0.56	1.04
H	1	0.63	0.95	0.47	1.12
	2	0.73	0.82	0.51	1.06
	3	0.83	0.65	0.58	0.96
	4	0.87	0.56	0.66	0.87
	5	0.91	0.48	0.72	0.75

[a]Abbreviations: PC, principal component; r^2, correlation coefficient; SDEC, standard deviation of errors in correlation; q^2, predictive correlation coefficient; SDEP, standard deviation of errors in prediction. *b* Models include the following variables: A, AMBER van der Waals and electrostatic interactions; B, AMBER van der Waals and DelPhi electrostatic interactions; C, AMBERn van der Waals, DelPhi electrostatic interactions, and ¢*G*desolv L ; D, AMBER van der Waals, DelPhi electrostatic interactions and ¢*G*desolv R ; E, AMBER van der Waals, DelPhi electrostatic interactions, and ¢*G*desolv Tyr442; F, AMBER van der Waals and electrostatic interactions including two water molecules;58 G, AMBER van der Waals and DelPhi electrostatic interactions including two water molecules; 58 H, AMBER van der Waals, DelPhi electrostatic interactions, ¢*G*desolv L , and ¢*G*desolv R .

According to the authors, the COMBINE model was externally validated when it was shown to predict the activity of an additional set of compounds that were not employed in model construction, see **Figure 12**.

Remarkably, the differences in inhibitory potency within the whole series were found to be finely tuned by the electrostatic contribution to the desolvation of the binding site and a network of secondary interactions established between the inhibitor and several protein residues that are distinct from those directly involved in the anchoring of the ligand. This information can now be used to advantage in the design of more potent inhibitors.

Figure 12. Superposition of the training set of inhibitors as found in their respective complexes with AChE. Relevant active site residues for one representative protein have been labeled. A semitransparent solvent-accessible surface envelops the side chains of the labeled residues (except Asp72) to delineate the active site cavity. Water molecules and all the hydrogen atoms have been omitted for clarity. Carbon atoms of tacrines, huprines, and dihydroquinazolines are shown in green, orange, and cyan, respectively.

1.14. Structure-based 3D QSAR and design of novel ACeH inhibitors

The paper Wolfgang Sippl [85] describes the construction, validation and application of a structure-based 3D QSAR model of novel acetylcholinesterase (AChE) inhibitors. Initial use was made of four X-ray structures of AChE complexed with small, non-specific inhibitors to create a model of the binding of recently developed aminopyridazine derivatives. Combined automated and manual docking methods were applied to dock the co-crystallized inhibitors into the binding pocket. Validation of the modelling process was achieved by comparing the predicted enzyme-bound conformationwith the known conformation in the X-ray structure. The successful prediction of the binding conformation of the known inhibitors gave confidence that we could use our model to evaluate the binding conformation of the aminopyridazine compounds. The alignment of 42 aminopyridazine compounds derived by the docking procedure was taken as the basis for a 3D QSAR analysis applying the GRID/GOLPE method. A model of high quality was obtained using the GRID water probe, as confirmed by the cross-validation method ($q^2_{LOO} = 0{:}937, q^2_{L50\%O} = 0{:}910$). The validated model, with information obtained from the calculated AChE-inhibitor complexes, was considered for the design of novel compounds. Seven designed inhibitors which were synthesized and tested were shown to be highly active. After performing our modelling study the X-ray structure of AChE complexed with donepezil, an inhibitor structurally related to the developed aminopyridazines. The good agreement found between the predicted binding conformation of the aminopyridazines and the one observed for donepezil in the crystal structure further supports our developed model.

2. New method for the study of new Acetylcholinesterase inhibitors.

2.1. Preface to new QSAR study Acetylcholinesterase inhibitors

Alzheimer´s disease (AD) [86] is a serious and degenerative disorder that causes a the gradual loss of neurons, and in spite of the efforts realized by the big pharmaceutical companies of the world, the origen of this pathology is still not very clear. Treatment of AD by ACh precursors and cholinergic agonists was ineffective or caused severe side effects. ACh hydrolysis by AChE causes termination of cholinergic neurotransmission. Therefore, compounds which inhibit AChE might significantly increase the levels of ACh depleted in AD. However, these drugs don't change the underlying disease process and may help only for a few months to a few years. We can see in this paper that the development of theoretical and QSAR models to study acetylcholinesterase inhibitors are usually not many achieved so far, and most of these works present docking studies. Watching this situation we need to develop QSAR models with acetylcholinesterase inhibitors. In this sense, quantitative structure-activity relationships (QSAR) could play an important role in studying these acetylcholinesterase inhibitors; QSARs can be used as predictive tools for the development of molecules [87, 88]. Computer-aided drug design techniques based on Quantitative Structure-Activity Relationships (QSAR) could play an important role in drug discovery programs. The QSAR approach involves the development of models that relate the structure of drugs with their biological activity against different targets [89, 90]. In principle, there are currently more than 1600 molecular descriptors that may be generalized and used to solve the problem outlined above [91]. Numerous different molecular descriptors have been reported to encode chemical structures in QSAR studies. Furthermore, there are multiple chemometric approaches that can, in principle, be selected for this step. Multiple linear regression (MLR) or linear discriminant analysis (LDA) [92], can be used to relate molecular structure (represented by molecular descriptors) with biological properties. In this article, we developed QSAR models for acetylcholinesterase inhibitors, LDA was constructed from more than 20 000 cases with more than 5 000 different molecules inhibitors of acetylcholinesterase obtained from ChEMBL database http://www.ebi.ac.uk/chembldb/index.php; in total we used more than 10 000 different molecules to develop the QSAR models. We used spectral moments molecular descriptors calculated with Modeslab software [71].

2.2. Methods

2.2.1. Linear classifier

A database from ChEMBL database [93] containing assayed AChE inhibitors was used (**Table SM** from the Supplementary Material, request to author). The Modeslab software[71] was utilized here and provides spectral moments descriptors[94-98]. The QSAR model was constructed with the multivariate regression technique, the LDA, employing the Forward stepwise method for the selection of variables. All statistical analyses and data exploration were carried out in STATISTICA 6.0[99] . In the actual work, the independent data test is used by splitting the data randomly in a training series used for a model construction and a cross-validation (CV) one. The general formula of the QSAR classification function is the following:

$$AChE_{score} = \sum W_m \cdot^m 2D_i + W_0 \qquad (1)$$

Where $AChE_{score}$ is the continuous and dimensionless score value for the $AChE_{score}$ /non- $AChE_{score}$ classification that gives relatively higher values to molecules with more probability to act as $AChE_{score}$, ^m2D_i are the 2*Ds* of type *m*, W_m is the coefficient (weights) of these indices in the QSAR model and W_0 is the independent term. The reported statistical parameters of the QSAR model are the following: N, χ^2, F, and p-level as well as Sensitivity, Specificity, and Accuracy for both training and CV . N is the number of molecules used to train the model, λ is Wilks statistic parameter, is Chi-square and p-level is the probability of error.

2.2.2. Data set

The data set used in this article was obtained from ChEMBL database. It has more than 20 000 cases and more than 5 000 different compounds inhibitors of AChE. In total we used more than 10 000 different molecules to develop the QSAR models obtained in ChEMBL. This is a database of bioactive drug-like small molecules, it contains 2D structures, calculated properties (e.g. logP, Molecular Weight, Lipinski Parameters, etc.) and abstracted bioactivities (e.g. binding constants, pharmacology and ADMET data). ChEMBL normalises the bioactivities into a uniform set of end-points and units where possible, and also tags the links between a molecular target and a published assay with a set of varying confidence levels. The data is abstracted and curated from the primary scientific literature, and covers a significant fraction of the SAR and discovery of modern drugs. The codes and activity for all compounds as well as the references used to collect them are depicted in **SM1** of the supplementary material file (request to author).

2.3. Results and Discussion

The present is a QSAR model for the probability of binding organic compounds to acetylcholinesterase receptor based only on the molecular connectivity of the drug and the protein receptor. Using this model we can predict the different relationships between the drug-protein connectivity same physicochemical property [100]. This work introduces a single linear QSAR equation model to classify drugs with acetylcholinesterase receptor. The best model found was:

$$AChE_{pred} = 3.56 \cdot \mu_0 - 0.21 \cdot \mu_2 + 15.47 \cdot \mu_3 - 5.3 \cdot \mu_4 - 8.96 \cdot \mu_5 + 0.76 \cdot \mu_6 + 1.11 \cdot \mu_7 + 2.11 \cdot \mu_8 + 0.69 \cdot \mu_{10} - 0.02 \cdot \mu_{11} + 0.002 \cdot \mu_{12} - 0.0002 \cdot \mu_{13} - 2.32x10^{-6} \cdot \mu_{15} - 8.6196 \qquad (2)$$

$$N = 22323 \quad \chi^2 = 10756.74 \quad \lambda = 0.3617 \quad p < 0.001$$

The nomenclature used in the descriptors of the equation is the same as establishing the Dragon software, where N is the number of compounds used for training, λ is the Wilks's statistic parameter, χ^2 is the Chi-square and p is the level of error. This model, with 13 variables, classifies correctly 1803 out of 2104 active (Sensitivity of 85.7%) and 12165 out of 12825 non-active (Specificity of 94.85%). Overall training Accuracy was 93.58%. The validation of the model was carried out by means of external predicting series. The model classifies correctly 921 out of 1053 active (87.45%) and 6040 out of 6341 non-active (95.25%) in validation series. Accuracy for validation series (predictability) was 94.18% (6961 out of 7394 DTPs). These results (**Table 3**) indicate that we developed an accurate model according to previous reports on the use of LDA in QSAR [101, 102].

Table 3. Results of the new LDA classification model.

Parameter	%	Class	active	No active
Analysis				
Sensitivity	94.85	No active	12165	660
Specificity	85.7	active	301	1803
Accuracy	93.58	Total		
Validation				
Sensitivity	95.25	n-active	6040	301
Specificity	87.45	active	132	921
Accuracy	94.18	Total		

CONCLUSIONS

Acetylcholinesterase functions and its implication in Alzheimer's disease have triggered an active search for potent and selective acetylcholinesterase inhibitors. Currently theoretical studies such as QSAR models have become a very useful tool in this context to substantially reduce time and resources consuming experiments. In this paper we can see that the development of theoretical and QSAR models to study acetylcholinesterase inhibitors are usually not many achieved so far, and most of these works present docking studies. Watching this situation we need to develop QSAR models with acetylcholinesterase inhibitors. In this sense, QSAR could play an important role in studying these acetylcholinesterase inhibitors. QSARs can be used as predictive tools for the development of molecules. In this work we developed a new LDA model using the ModesLab descriptors, based on a large database using about 10,000 different drugs obtained from the ChemBL server.

Acknowledgements

Prado-Prado F. thanks sponsorships for research position at the University of Santiago de Compostela from *Angeles Alvariño*, Xunta de Galicia.

REFERENCES.

[1] Querfurth HW, LaFerla FM. Alzheimer's disease. *N Engl J Med* **2010**; 362(4): 329-44.

[2] Goate A, Chartier-Harlin MC, Mullan M, Brown J, Crawford F, Fidani L, Giuffra L, Haynes A, Irving N, James L, et al. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* **1991**; 349(6311): 704-6.

[3] Levy-Lahad E, Wasco W, Poorkaj P, Romano DM, Oshima J, Pettingell WH, Yu CE, Jondro PD, Schmidt SD, Wang K, et al. Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science* **1995**; 269(5226): 973-7.

[4] Sherrington R, Rogaev EI, Liang Y, Rogaeva EA, Levesque G, Ikeda M, Chi H, Lin C, Li G, Holman K, et al. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* **1995**; 375(6534): 754-60.

[5] Citron M, Westaway D, Xia W, Carlson G, Diehl T, Levesque G, Johnson-Wood K, Lee M, Seubert P, Davis A, Kholodenko D, Motter R, Sherrington R, Perry B, Yao H, Strome R, Lieberburg I, Rommens J, Kim S, Schenk D, Fraser P, St George Hyslop P, Selkoe DJ. Mutant presenilins of Alzheimer's disease increase production of 42-residue amyloid beta-protein in both transfected cells and transgenic mice. *Nat Med* **1997**; 3(1): 67-72.

[6] Holcomb L, Gordon MN, McGowan E, Yu X, Benkovic S, Jantzen P, Wright K, Saad I, Mueller R, Morgan D, Sanders S, Zehr C, O'Campo K, Hardy J, Prada CM, Eckman C, Younkin S, Hsiao K, Duff K. Accelerated Alzheimer-type phenotype in transgenic mice carrying both mutant amyloid precursor protein and presenilin 1 transgenes. *Nat Med* **1998**; 4(1): 97-100.

[7] Cole SL, Vassar R. The Alzheimer's disease beta-secretase enzyme, BACE1. *Mol Neurodegener* **2007**; 2: 22.

[8] Cole SL, Vassar R. The Basic Biology of BACE1: A Key Therapeutic Target for Alzheimer's Disease. *Curr Genomics* **2007**; 8(8): 509-30.

[9] Sussman JL, Harel M, Frolow F, Oefner C, Goldman A, Toker L, Silman I. Atomic structure of acetylcholinesterase from Torpedo californica: a prototypic acetylcholine-binding protein. *Science* **1991**; 253(5022): 872-9.

[10] Botti SA, Felder CE, Lifson S, Sussman JL, Silman I. A modular treatment of molecular traffic through the active site of cholinesterase. *Biophys J* **1999**; 77(5): 2430-50.

[11] Chou KC. Structural bioinformatics and its impact to biomedical science. *Curr Med Chem* **2004**; 11(16): 2105-34.

[12] Chou KC, Wei DQ, Du QS, Sirois S, Zhong WZ. Progress in computational approach to drug development against SARS. *Curr Med Chem* **2006**; 13(27): 3263-70.

[13] Todeschini R, Consonni V. Handbook of Molecular Descriptors. Wiley-VCH: **2002**.

[14] Estrada E, Uriarte E. Recent advances on the role of topological indices in drug discovery research. *Curr Med Chem* **2001**; 8: 1573-88.

[15] González-Díaz H, González-Díaz Y, Santana L, Ubeira FM, Uriarte E. Proteomics, networks and connectivity indices. *Proteomics* **2008**; 8: 750-78.

[16] Nunez MB, Maguna FP, Okulik NB, Castro EA. QSAR modeling of the MAO inhibitory activity of xanthones derivatives. *Bioorg Med Chem Lett* **2004**; 14(22): 5611-7.

[17] Terada M, Inaba M, Yano Y, Hasuma T, Nishizawa Y, Morii H, Otani S. Growth-inhibitory effect of a high glucose concentration on osteoblast-like cells. *Bone* **1998**; 22(1): 17-23.

[18] Santana L, Uriarte E, González-Díaz H, Zagotto G, Soto-Otero R, Mendez-Alvarez E. A QSAR model for in silico screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins. *J Med Chem* **2006**; 49(3): 1149-56.

[19] Marrero-Ponce Y. Linear indices of the "molecular pseudograph's atom adjacency matrix": definition, significance-interpretation, and application to QSAR analysis of flavone derivatives as HIV-1 integrase inhibitors. *J Chem Inf Comput Sci* **2004**; 44(6): 2010-26.

[20] Vilar S, Santana L, Uriarte E. Probabilistic neural network model for the in silico evaluation of anti-HIV activity and mechanism of action. *J Med Chem* **2006**; 49(3): 1118-24.

[21] Marrero-Ponce Y, Khan MT, Casanola Martin GM, Ather A, Sultankhodzhaev MN, Torrens F, Rotondo R. Prediction of Tyrosinase Inhibition Activity Using Atom-Based Bilinear Indices. *ChemMedChem* **2007**; 2(4): 449-78.

[22] Casanola-Martin GM, Marrero-Ponce Y, Khan MT, Ather A, Sultan S, Torrens F, Rotondo R. TOMOCOMD-CARDD descriptors-based virtual screening of tyrosinase inhibitors: evaluation of different classification model combinations using bond-based linear indices. *Bioorg Med Chem* **2007**; 15(3): 1483-503.

[23] Casanola-Martin GM, Marrero-Ponce Y, Khan MT, Ather A, Khan KM, Torrens F, Rotondo R. Dragon method for finding novel tyrosinase inhibitors: Biosilico identification and experimental in vitro assays. *Eur J Med Chem* **2007**; 42(11-12): 1370-81.

[24] Chen J, Shen B. Computational Analysis of Amino Acid Mutation: a Proteome Wide Perspective. *Curr Proteomics* **2009**; 6(4): 228-34.
[25] Chou KC. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr Proteomics* **2009**; 6(4): 262-74.
[26] Berca MN, Duardo-Sanchez A, González-Díaz H, Pazos A, Munteanu CR. In: González-Díaz H, Prado-Prado FJ, García-Mera X, Eds. Complex Network Entropy: From Molecules to Biology, Parasitology, Technology, Social, Legal, and Neurosciences. Kerala, India, Transworld Research Network **2011**;127-42.
[27] Vilar S, Gonzalez-Diaz H, Santana L, Uriarte E. A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer. *J Theor Biol* **2009**; 261(3): 449-58.
[28] Concu R, Dea-Ayuela MA, Perez-Montoto LG, Prado-Prado FJ, Uriarte E, Bolas-Fernandez F, Podda G, Pazos A, Munteanu CR, Ubeira FM, Gonzalez-Diaz H. 3D entropy and moments prediction of enzyme classes and experimental-theoretic study of peptide fingerprints in Leishmania parasites. *Biochim Biophys Acta* **2009**; 1794(12): 1784-94.
[29] Fjodorova N, Novich M, Vrachko M, Kharchevnikova N, Zholdakova Z, Sinitsyna O, Benfenati E. Regulatory assessment of chemicals within OECD member countries, EU and in Russia. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev* **2008**; 26(1): 40-88.
[30] Duardo-Sanchez A. In: González-Díaz H, Prado-Prado FJ, García-Mera X, Eds. Complex Network Entropy: From Molecules to Biology, Parasitology, Technology, Social, Legal, and Neurosciences. Kerala, India, Transworld Research Network **2011**;107-14.
[31] Gonzalez-Diaz H. Quantitative studies on Structure-Activity and Structure-Property Relationships (QSAR/QSPR). *Curr Top Med Chem* **2008**; 8(18): 1554.
[32] Ivanciuc O. Weka machine learning for predicting the phospholipidosis inducing potential. *Curr Top Med Chem* **2008**; 8(18): 1691-709.
[33] Gonzalez-Diaz H, Prado-Prado F, Ubeira FM. Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. *Curr Top Med Chem* **2008**; 8(18): 1676-90.
[34] Duardo-Sanchez A, Patlewicz G, Lopez-Diaz A. Current topics on software use in medicinal chemistry: intellectual property, taxes, and regulatory issues. *Curr Top Med Chem* **2008**; 8(18): 1666-75.
[35] Wang JF, Wei DQ, Chou KC. Drug candidates from traditional chinese medicines. *Curr Top Med Chem* **2008**; 8(18): 1656-65.
[36] Pillay D, Cane PA, Ratcliffe D, Atkins M, Cooper D. Evolution of lamivudine-resistant hepatitis B virus and HIV-1 in co-infected individuals: an analysis of the CAESAR study. CAESAR co-ordinating committee. *AIDS* **2000**; 14(9): 1111-6.
[37] Gonzalez MP, Teran C, Saiz-Urra L, Teijeira M. Variable selection methods in QSAR: an overview. *Curr Top Med Chem* **2008**; 8(18): 1606-27.
[38] Caballero J, Fernandez M. Artificial neural networks from MATLAB in medicinal chemistry. Bayesian-regularized genetic neural networks (BRGNN): application to the prediction of the antagonistic activity against human platelet thrombin receptor (PAR-1). *Curr Top Med Chem* **2008**; 8(18): 1580-605.
[39] Wang JF, Wei DQ, Chou KC. Pharmacogenomics and personalized use of drugs. *Curr Top Med Chem* **2008**; 8(18): 1573-9.
[40] Vilar S, Cozza G, Moro S. Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. *Curr Top Med Chem* **2008**; 8(18): 1555-72.
[41] Prado-Prado FJ, Garcia-Mera X, Gonzalez-Diaz H. Multi-target spectral moment QSAR versus ANN for antiparasitic drugs against different parasite species. *Bioorg Med Chem* **2010**; 18(6): 2225-31.
[42] Lin YY, Qi Y, Lu JY, Pan X, Yuan DS, Zhao Y, Bader JS, Boeke JD. A comprehensive synthetic genetic interaction network governing yeast histone acetylation and deacetylation. *Genes Dev* **2008**; 22(15): 2062-74.
[43] Zotenko E, Mestre J, O'Leary DP, Przytycka TM. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput Biol* **2008**; 4(8): e1000140.
[44] Hase T, Niimura Y, Kaminuma T, Tanaka H. Non-uniform survival rate of heterodimerization links in the evolution of the yeast protein-protein interaction network. *PLoS ONE* **2008**; 3(2): e1667.
[45] Kafri R, Dahan O, Levy J, Pilpel Y. Preferential protection of protein interaction network hubs in yeast: evolved functionality of genetic redundancy. *Proc Natl Acad Sci U S A* **2008**; 105(4): 1243-8.
[46] Jin G, Zhang S, Zhang XS, Chen L. Hubs with network motifs organize modularity dynamically in the protein-protein interaction network of yeast. *PLoS ONE* **2007**; 2(11): e1207.
[47] Fernandez A. Molecular basis for evolving modularity in the yeast protein interaction network. *PLoS Comput Biol* **2007**; 3(11): e226.
[48] Batada NN, Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hurst LD, Tyers M. Stratus not altocumulus: a new view of the yeast protein interaction network. *PLoS Biol* **2006**; 4(10): e317.
[49] Li L, Huang Y, Xia X, Sun Z. Preferential duplication in the sparse part of yeast protein interaction network. *Mol Biol Evol* **2006**; 23(12): 2467-73.
[50] Chou KC. Graphic rule for drug metabolism systems. *Curr Drug Metab* 11(4): 369-78.

[51] Nassif H, Al-Ali H, Khuri S, Keirouz W. Prediction of protein-glucose binding sites using support vector machines. *Proteins* **2009**; 77(1): 121-32.
[52] Del Rio A, Baldi BF, Rastelli G. Activity prediction and structural insights of extracellular signal-regulated kinase 2 inhibitors with molecular dynamics simulations. *Chemical biology & drug design* **2009**; 74(6): 630-5.
[53] Garcia I, Fall Y, Gomez G. QSAR, docking, and CoMFA studies of GSK3 inhibitors. *Curr Pharm Des* **2010**; 16(24): 2666-75.
[54] Le T, Tseng TT, Saier MH, Jr. Flexible programs for the prediction of average amphipathicity of multiply aligned homologous proteins: application to integral membrane transport proteins. *Mol Membr Biol* **1999**; 16(2): 173-9.
[55] Gonzalez-Diaz H, Romaris F, Duardo-Sanchez A, Perez-Mototo LG, Prado-Prado F, Patlewicz G, Ubeira FM. Predicting drugs and proteins in parasite infections with topological indices of complex networks: theoretical backgrounds, aplications, and legal issues. *Curr Pharm Des* **2010**; 16(24): 2737-64.
[56] Marrero-Ponce Y, Casanola-Martin GM, Khan MT, Torrens F, Rescigno A, Abad C. Ligand-Based Computer-Aided Discovery of Tyrosinase Inhibitors. Applications of the TOMOCOMD-CARDD Method to the Elucidation of New Compounds. *Curr Pharm Des* **2010**; 16(24): 2601-24.
[57] Munteanu CR, Fernandez-Blanco E, Seoane JA, Izquierdo-Novo P, Rodriguez-Fernandez JA, Prieto-Gonzalez JM, Rabunal JR, Pazos A. Drug discovery and design for complex diseases through QSAR computational methods. *Curr Pharm Des* **2010**; 16(24): 2640-55.
[58] Roy K, Ghosh G. Exploring QSARs with Extended Topochemical Atom (ETA) Indices for Modeling Chemical and Drug Toxicity. *Curr Pharm Des* **2010**; 16(24): 2625-39.
[59] Speck-Planche A, Scotti MT, de Paulo-Emerenciano V. Current pharmaceutical design of antituberculosis drugs: future perspectives. *Curr Pharm Des* **2010**; 16(24): 2656-65.
[60] Veluraja K, Seethalakshmi AN. Dynamics of sialyl Lewis(a) in aqueous solution and prediction of the structure of the sialyl Lewis(a)-SelectinE complex. *J Theor Biol* **2008**; 252(1): 15-23.
[61] Bhattacharjee B, Jayadeepa RM, Banerjee S, Joshi J, Middha SK, Mole JP, Samuel J. Review of Complex Network and Gene Ontology in pharmacology approaches: Mapping natural compounds on potential drug target Colon Cancer network. *Current Bioinformatics* **2011**; 6(1): 44-52.
[62] Chiş O, Dumitru O, Concu R, Shen B. Reviewing Yeast Network and report of new Stochastic-Credibility cell cycle models. *Current Bioinformatics* **2011**; 6(1): 35-43.
[63] Dave K, Banerjee A. Bioinformatics analysis of functional relations between CNPs regions. *Current Bioinformatics* **2011**; 6(1): 122-8.
[64] Duardo-Sanchez A, Patlewicz G, González-Díaz H. A Review of Network Topological Indices from Chem-Bioinformatics to Legal Sciences and back. *Current Bioinformatics* **2011**; 6(11): 53-70.
[65] García I, Fall Y, Gómez G. Trends in Bioinformatics and Chemoinformatics of Vitamin D analogues and their protein targets. *Current Bioinformatics* **2011**; 6(1): 16-24.
[66] Ivanciuc T, Ivanciuc O, Klein DJ. Network-QSAR with Reaction Poset Quantitative Superstructure-Activity Relationships (QSSAR) for PCB Chromatographic Properties. *Current Bioinformatics* **2011**; 6(1): 25-34.
[67] Prado-Prado F, Escobar-Cubiella M, García-Mera X. Review of Bioinformatics and QSAR studies of β-secretase inhibitors. *Current Bioinformatics* **2011**; 6(1): 3-15.
[68] Riera-Fernández P, Munteanu CR, Pedreira-Souto N, Duardo-Sanchez A, González-Díaz H. Definition of Markov-Harary Invariants and Review of Classic Topological Indices and Databases of Biology, Parasitology, Technology, and Social-Legal Networks. *Current Bioinformatics* **2011**; 6(1): 94-121.
[69] Speck-Planche A, Cordeiro MNDS. Application of Bioinformatics for the search of novel anti-viral therapies: Rational design of anti-herpes agents. *Current Bioinformatics* **2011**; 6(1): 81-93.
[70] Wan SB, Hu LL, Niu S, Wang K, Cai YD, Lu WC, Chou KC. Identification of multiple subcellular locations for proteins in budding yeast. *Current Bioinformatics* **2011**; 6(1): 71-80.
[71] Estrada E. On the topological sub-structural molecular design (TOSS-MODE) in QSPR/QSAR and drug design research. *SAR QSAR Environ Res* **2000**; 11(1): 55-73.
[72] Brady CP, Dowd AJ, Tort J, Roche L, Condon B, O'Neill SM, Brindley PJ, Dalton JP. The cathepsin L-like proteinases of liver fluke and blood fluke parasites of the trematode genera Fasciola and Schistosoma. *Biochem Soc Trans* **1999**; 27(4): 740-5.
[73] Badawy ME, El-Arami SA, Abdelgaleil SA. Acaricidal and quantitative structure activity relationship of monoterpenes against the two-spotted spider mite, Tetranychus urticae. *Exp Appl Acarol* **2010**; 52(3): 261-74.
[74] Vitorovic-Todorovic MD, Juranic IO, Mandic LM, Drakulic BJ. 4-Aryl-4-oxo-N-phenyl-2-aminylbutyramides as acetyl- and butyrylcholinesterase inhibitors. Preparation, anticholinesterase activity, docking study, and 3D structure-activity relationship based on molecular interaction fields. *Bioorganic & medicinal chemistry* **2010**; 18(3): 1181-93.

[75] Lv W, Xue Y. Prediction of acetylcholinesterase inhibitors and characterization of correlative molecular descriptors by machine learning methods. *Eur J Med Chem* **2010**; 45(3): 1167-72.

[76] Ul-Haq Z, Mahmood U, Jehangir B. Ligand-based 3D-QSAR studies of physostigmine analogues as acetylcholinesterase inhibitors. *Chemical biology & drug design* **2009**; 74(6): 571-81.

[77] Chaudhaery SS, Roy KK, Saxena AK. Consensus superiority of the pharmacophore-based alignment, over maximum common substructure (MCS): 3D-QSAR studies on carbamates as acetylcholinesterase inhibitors. *Journal of chemical information and modeling* **2009**; 49(6): 1590-601.

[78] Armutlu P, Ozdemir ME, Uney-Yuksektepe F, Kavakli IH, Turkay M. Classification of drug molecules considering their IC50 values using mixed-integer linear programming based hyper-boxes method. *BMC Bioinformatics* **2008**; 9: 411.

[79] Shen LL, Liu GX, Tang Y. Molecular docking and 3D-QSAR studies of 2-substituted 1-indanone derivatives as acetylcholinesterase inhibitors. *Acta pharmacologica Sinica* **2007**; 28(12): 2053-63.

[80] Ahn JH, Shin MS, Jung SH, Kim JA, Kim HM, Kim SH, Kang SK, Kim KR, Rhee SD, Park SD, Lee JM, Lee JH, Cheon HG, Kim SS. Synthesis and structure-activity relationship of novel indene N-oxide derivatives as potent peroxisome proliferator activated receptor gamma (PPARgamma) agonists. *Bioorg Med Chem Lett* **2007**; 17(18): 5239-44.

[81] Akula N, Lecanu L, Greeson J, Papadopoulos V. 3D QSAR studies of AChE inhibitors based on molecular docking scores and CoMFA. *Bioorganic & medicinal chemistry letters* **2006**; 16(24): 6277-80.

[82] Fontaine F, Pastor M, Zamora I, Sanz F. Anchor-GRIND: filling the gap between standard 3D QSAR and the GRid-INdependent descriptors. *J Med Chem* **2005**; 48(7): 2687-94.

[83] Guo J, Hurley MM, Wright JB, Lushington GH. A docking score function for estimating ligand-protein interactions: application to acetylcholinesterase inhibition. *J Med Chem* **2004**; 47(22): 5492-500.

[84] Martin-Santamaria S, Munoz-Muriedas J, Luque FJ, Gago F. Modulation of binding strength in several classes of active site inhibitors of acetylcholinesterase studied by comparative binding energy analysis. *J Med Chem* **2004**; 47(18): 4471-82.

[85] Sippl W, Contreras JM, Parrot I, Rival YM, Wermuth CG. Structure-based 3D QSAR and design of novel acetylcholinesterase inhibitors. *J Comput Aided Mol Des* **2001**; 15(5): 395-410.

[86] Salmon SA, Watts JL. Minimum inhibitory concentration determinations for various antimicrobial agents against 1570 bacterial isolates from turkey poults. *Avian Dis* **2000**; 44(1): 85-98.

[87] Chou KC. Review: Structural bioinformatics and its impact to biomedical science. *Curr Med Chem* **2004**; 11: 2105-34.

[88] Chou KC, D. Q. Wei, Q. S. Du, S. Sirois, and W. Z. Zhong. . Review: Progress in computational approach to drug development against SARS. *Curr Med Chem* **2006**; 13: 3263-70.

[89] Prado-Prado FJ, Gonzalez-Diaz H, de la Vega OM, Ubeira FM, Chou KC. Unified QSAR approach to antimicrobials. Part 3: first multi-tasking QSAR model for input-coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. *Bioorg Med Chem* **2008**; 16(11): 5871-80.

[90] Prado-Prado FJ, de la Vega OM, Uriarte E, Ubeira FM, Chou KC, Gonzalez-Diaz H. Unified QSAR approach to antimicrobials. 4. Multi-target QSAR modeling and comparative multi-distance study of the giant components of antiviral drug-drug complex networks. *Bioorg Med Chem* **2009**; 17: 569–75.

[91] Kubinyi H. Quantitative structure-activity relationships (QSAR) and molecular modelling in cancer research. *J Cancer Res Clin Oncol* **1990**; 116(6): 529-37.

[92] Prado-Prado FJ, Borges F, Perez-Montoto LG, Gonzalez-Diaz H. Multi-target spectral moment: QSAR for antifungal drugs vs. different fungi species. *Eur J Med Chem* **2009**.

[93] Overington J. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. *J Comput Aided Mol Des* **2009**; 23(4): 195-8.

[94] Estrada E, Patlewicz G, Chamberlain M, Basketter D, Larbey S. Computer-aided knowledge generation for understanding skin sensitization mechanisms: the TOPS-MODE approach. *Chem Res Toxicol* **2003**; 16(10): 1226-35.

[95] Estrada E, Uriarte E, Gutierrez Y, Gonzalez H. Quantitative structure-toxicity relationships using TOPS-MODE. 3. Structural factors influencing the permeability of commercial solvents through living human skin. *SAR QSAR Environ Res* **2003**; 14(2): 145-63.

[96] Estrada E, Gonzalez H. What are the limits of applicability for graph theoretic descriptors in QSPR/QSAR? Modeling dipole moments of aromatic compounds with TOPS-MODE descriptors. *J Chem Inf Comput Sci* **2003**; 43(1): 75-84.

[97] Estrada E, Vilar S, Uriarte E, Gutierrez Y. In silico studies toward the discovery of new anti-HIV nucleoside compounds with the use of TOPS-MODE and 2D/3D connectivity indices. 1. Pyrimidyl derivatives. *J Chem Inf Comput Sci* **2002**; 42(5): 1194-203.

[98] Estrada E. Quantum-chemical foundations of the topological substructural molecular design. *J Phys Chem A* **2008**; 112(23): 5208-17.

[99] Yoshii F, Hirono S. Construction of a quantitative three-dimensional model for odor quality using comparative molecular field analysis (CoMFA). *Chem Senses* **1996**; 21(2): 201-10.

[100] Gonzalez-Diaz H, Saiz-Urra L, Molina R, Santana L, Uriarte E. A model for the recognition of protein kinases based on the entropy of 3D van der Waals interactions. *Journal of proteome research* **2007**; 6(2): 904-8.

[101] Alvarez-Ginarte YM, Marrero-Ponce Y, Ruiz-Garcia JA, Montero-Cabrera LA, Vega JM, Noheda Marin P, Crespo-Otero R, Zaragoza FT, Garcia-Domenech R. Applying pattern recognition methods plus quantum and physico-chemical molecular descriptors to analyze the anabolic activity of structurally diverse steroids. *J Comput Chem* **2007**.

[102] Morales AH, Rodríguez-Borges JE, García-Mera X, Fernández F, Dias-Sueiro-Cordeiro MN. Probing the Anticancer Activity of Nucleoside Analogues: A QSAR Model Approach Using an Internally Consistent Training Set. *J Med Chem* **2007**; 50: 1537-45.

European Journal of Medicinal Chemistry 46 (2011) 5838–5851

Contents lists available at SciVerse ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

2D MI-DRAGON: A new predictor for protein–ligands interactions and theoretic-experimental studies of US FDA drug-target network, oxoisoaporphine inhibitors for MAO-A and human parasite proteins

Francisco Prado-Prado [a,*], Xerardo García-Mera [a], Manuel Escobar [a], Eduardo Sobarzo-Sánchez [b], Matilde Yañez [c], Pablo Riera-Fernandez [d], Humberto González-Díaz [d]

[a] Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela (USC) 15782, Spain
[b] Department of Pharmaceutical Technology, Faculty of Pharmacy, USC 5782, Spain
[c] Department of Pharmacology, Faculty of Pharmacy, USC 15782, Spain
[d] Department of Microbiology and Parasitology, Faculty of Pharmacy, USC 15782, Spain

ARTICLE INFO

Article history:
Received 20 June 2011
Received in revised form
22 September 2011
Accepted 26 September 2011
Available online 1 October 2011

Keywords:
Drug–Protein interaction complex networks
Protein structure networks
Multi-target QSAR
Markov model
MAO A inhibitors

ABSTRACT

There are many pairs of possible Drug–Proteins Interactions that may take place or not (DPIs/nDPIs) between drugs with high affinity/non-affinity for different proteins. This fact makes expensive in terms of time and resources, for instance, the determination of all possible ligands–protein interactions for a single drug. In this sense, we can use Quantitative Structure–Activity Relationships (QSAR) models to carry out rational DPIs prediction. Unfortunately, almost all QSAR models predict activity against only one target. To solve this problem we can develop multi-target QSAR (mt-QSAR) models. In this work, we introduce the technique 2D MI-DRAGON a new predictor for DPIs based on two different well-known software. We use the software MARCH-INSIDE (MI) to calculate 3D structural parameters for targets and the software DRAGON was used to calculated 2D molecular descriptors all drugs showing known DPIs present in the Drug Bank (US FDA benchmark dataset). Both classes of parameters were used as input of different Artificial Neural Network (ANN) algorithms to seek an accurate non-linear mt-QSAR predictor. The best ANN model found is a Multi-Layer Perceptron (MLP) with profile MLP 21:21-31-1:1. This MLP classifies correctly 303 out of 339 DPIs (Sensitivity = 89.38%) and 480 out of 510 nDPIs (Specificity = 94.12%), corresponding to training Accuracy = 92.23%. The validation of the model was carried out by means of external predicting series with Sensitivity = 92.18% (625/678 DPIs; Specificity = 90.12% (730/780 nDPIs) and Accuracy = 91.06%. 2D MI-DRAGON offers a good opportunity for fast-track calculation of all possible DPIs of one drug enabling us to re-construct large drug-target or DPIs Complex Networks (CNs). For instance, we reconstructed the CN of the US FDA benchmark dataset with 855 nodes 519 drugs + 336 targets). We predicted CN with similar topology (observed and predicted values of average distance are equal to 6.7 *vs.* 6.6). These CNs can be used to explore large DPIs databases in order to discover both new drugs and/or targets. Finally, we illustrated in one theoretic-experimental study the practical use of 2D MI-DRAGON. We reported the prediction, synthesis, and pharmacological assay of 10 different oxoisoaporphines with MAO-A inhibitory activity. The more active compound OXO5 presented IC_{50} = 0.00083 μM, notably better than the control drug Clorgyline.

1. Introduction

The prediction of interactions between organic compounds to form drug-target pairs (DPIs) is a source piece on the combination of bioinformatics and proteome research towards drug discovery. Therefore, there is a strong incentive to develop new methods capable of detecting these potential drug-target interactions efficiently [1]. In this sense, we can use Quantitative Structure–Activity Relationships (QSAR) models [2] to carry out rational DPIs prediction. However, the use of QSAR techniques to predict DPIs is an area less explored until now. Classic QSAR models are equations that connect the structure of the drug, expressed by means of molecular descriptors, with the biological function. In classic QSAR studies we can use different free or commercially

* Corresponding author. Faculty of Pharmacy, University of Santiago de Compostela, 15782, Spain. Fax: +34 981 594912.
E-mail address: francisco.prado@usc.e (F. Prado-Prado).

0223-5234/$ – see front matter © 2011 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2011.09.045

available software to calculate structural parameters of the drug. Some of the more known software we can use to reach this goal are: DRAGON, CODESSA [3], MODES-LAB [4], TOMO-COMD [5], and MARCH-INSIDE (MI) [6]. The software DRAGON is one of the more complete calculating more than 1600 descriptors for drug structure including as zero- (0D) one- (1D), two- (2D), three-dimensional (3D) parameters. Unfortunately almost all QSAR models are able to predict the activity of drugs against only one target. To solve this problem we can develop multi-target QSAR (mt-QSAR) models to predict DPIs [7]. One way to develop this class mt-QSAR is incorporating into the QSAR equation parameters of the structure of the target (protein, DNA, RNA, etc.) in addition to the structural parameters of the drug present in classic QSAR. Ideally, we should use structural parameters of both drug and target calculated with the same software in order to standardize the entire algorithm. Disappointedly, many of known software used for QSAR have been developed to calculate only molecular indices of drugs. In any case, the theory used in these software packages can be easily extended to calculate molecular descriptors of targets. Consequently, it is only a matter of time that the authors of these programs develop new upgrades incorporating also molecular descriptors of the targets. That is the case of TOMO-COMD and MI; both have incorporated the calculation of drug and target structural parameters. In any case, until the best of our knowledge only MI has been used to develop and publish

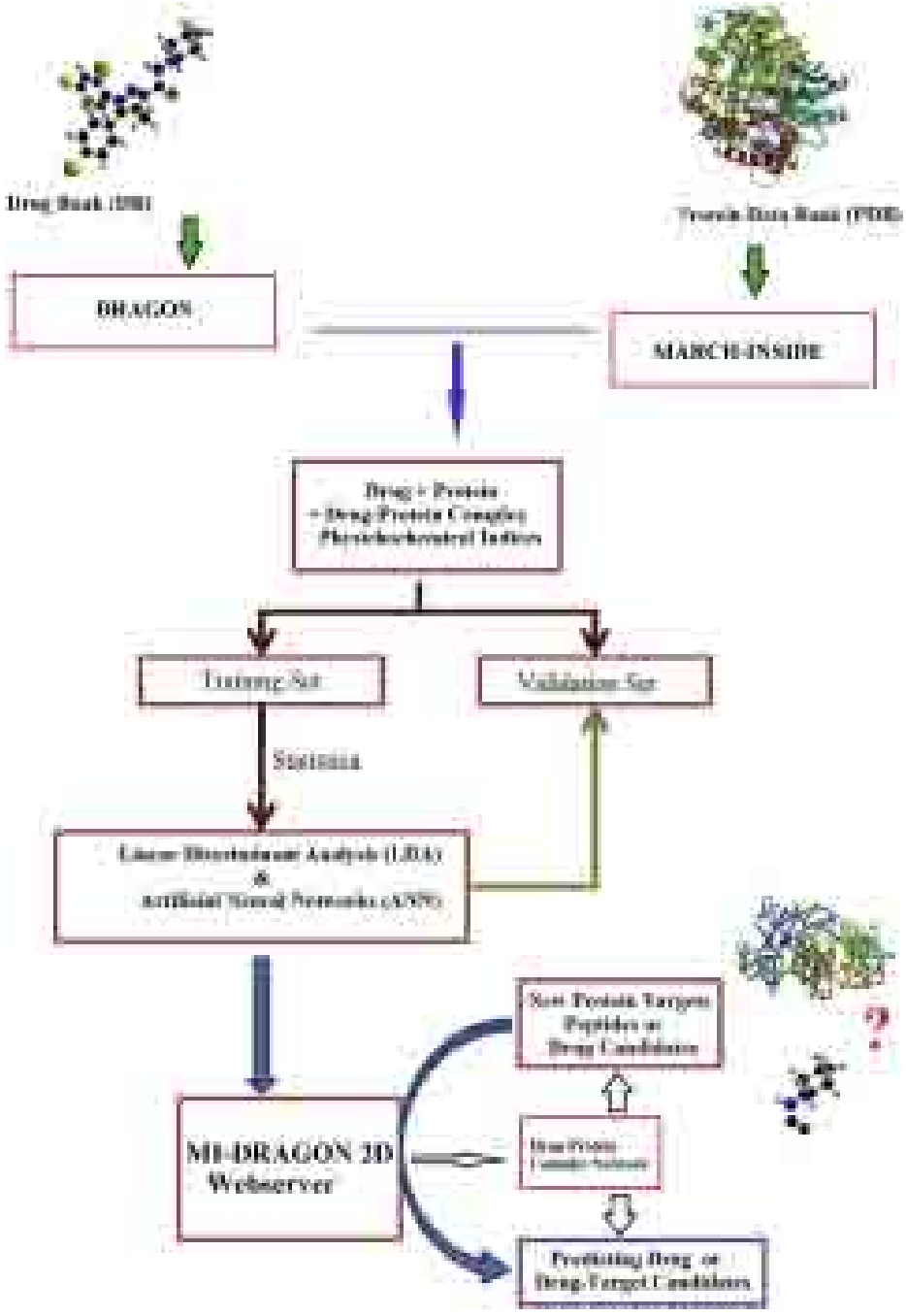


Fig. 1. Flowchart of all steps given in this work to develop the new model.

a QSAR predictor for DPIs. For instance, very recently we developed MIND-BEST [8] and NL MIND-BEST [9] two predictors based on structural parameters of drugs and proteins calculated with software MI.

As was mentioned in the previous paragraph we can seek a QSAR predictor for DPIs using molecular descriptors of both drug and target. Nevertheless, many of the aforementioned software can calculate only structural parameters for drugs and only a few of them have been upgraded to incorporate parameters of targets. In this sense, we propose to seek QSAR models to predict DPIs using two different software packages (one for drug parameters and other for target ones) as an alternative to the yet limited use of one program only. In a last step, we can use this new type of QSAR model for the prediction of all possible DPIs/nDPIs (non-drug-protein interaction) in the global set of relationships between protein targets and all drugs to form the complex network (CN) of all DPIs. This type of CN for DPIs may become of the major relevance for the discovery of new drugs and targets as well. For instance, Yildirim, et al. [10] have built a bipartite graph composed of US Food and Drug Administration (US FDA) approved drugs and proteins linked by drug-target binary associations. The resulting network connects most drugs into a highly interlinked giant component, with strong local clustering of drugs of similar types according to Anatomical Therapeutic Chemical classification. Topological analyses of this network quantitatively showed an overabundance of 'follow-on' drugs, that is, drugs that target already targeted proteins.

In this work, we developed for first time one mt-QSAR predictor of DPIs called 2D MI-DRAGON. The new technique combines the software DRAGON [11] to calculate structural parameters of drugs with software MI to calculate structural parameters of the target. The technique 2D MI-DRAGON uses these descriptors as input of one Artificial Neural Network (ANN) to seek the model. Both training and validation of the model was carried out by means of learning and external predicting series containing structural parameters for all DPIs present in the Drug Bank (US FDA benchmark dataset) downloaded from Drug Bank [12–15]. We also compare this model with other ANN models developed in this work and Machine Learning (ML) classifiers published before to address the same problem. A very good MI-DRAGON QSAR model was obtained, and the subsequent combined QSAR & CN analysis may become of major importance for the prediction of the activity of new compounds against different targets or the discovery of new targets. In this sense we reported an illustrative study that combines both experiment and theory to show how to use this model in practical situations. We reported the prediction, synthesis, and pharmacological assay of oxoisoaporphines with MAO-A inhibitory activity. In Fig. 1 we depict a flowchart with the main steps given in this work to train and validate the ANN classifier (Fig. 2).

2. Materials and methods

2.1. Computational methods

2.1.1. MI-DRAGON technique

2.1.1.1. Parameters for drugs. The DRAGON software 4.0 [11] was utilized here to calculate the parameters of drugs. This software provides near to 1600 descriptors classified as zero- (0D) one- (1D), two- (2D), and three-dimensional (3D) descriptors. It depends on whether they are computed from the chemical formula, substructure list representation, molecular graph or geometrical representation of the molecule, respectively [16,17]. In this work, we calculated only 0D – to – 2D descriptors. We use these descriptors because this way the drugs not optimized for use with 3D descriptors. We used specifically the following descriptors: 2D autocorrelations, Burden eigenvalues, topological charge indices,

Fig. 2. Snapshot of LOMETS server used to predict 3D structure of peptides.

eigenvalue-based indices, functional group counts, atoms-centered fragments, charge descriptors and molecular properties.

2.1.1.2. Parameter for targets (proteins). In previous works, we have predicted protein function based on 3D parameters of protein structure calculated with software MI [18,19]. In this paper, we used specifically the 3D-electrostatic potential parameters ξ_k. These values were used as inputs to construct the QSAR model together with the structural parameters of drugs. The detailed explanation of the procedure has been published before. Therefore, we provide herein only the more general formula for these potentials and some general explanations [20]:

$$\xi_m = \xi_k(R) = \sum_{j=1\in R}^{n} p_k(j)\cdot\xi(j) \quad (1)$$

The average general potentials depend on the absolute probabilities $p_k(j)$ and the total potential with which the aminoacid j-th interact with the rest of aminoacids. These are the probabilities with which the amino acids interact with other amino acids placed at a distance equals to k-times the cut-off distance ($r_{ij} = k\cdot r_{\text{cut-off}}$). The method uses an MCM to calculate these probabilities; which also depend on the 3D interactions between all pairs of aminoacids placed at distance r_{ij} in $\boldsymbol{r}_3$ in the protein structure. However, for the sake of simplicity, a truncation or cut-off function α_{ij} is applied in such a way that a short-term interaction takes place in a first approximation only between neighboring aa ($\alpha_{ij} = 1$ if $r_{ij} < r_{\text{cut-off}}$). Otherwise, the interaction is banished ($\alpha_{ij} = 0$). The relationship α_{ij} may be visualized in the form of a protein structure complex network. In this network the nodes are the C_α atoms of the aminoacids and the edges connect pairs of aminoacids with $\alpha_{ij} = 1$. This network can be understood in terms of aminoacid–aminoacid protein contact maps [21] in Euclidean 3D space $\boldsymbol{r}_3 = (x, y, z)$ coordinates of the C_α atoms of aminoacids listed on protein PDB files. In recent works we published different examples of these networks [22,23]. For the purposes of the calculation, all water molecules and metal ions were removed [20]. All calculations were carried out with our in-house software MI [20]. For calculation the MI software always uses the full matrix, never a sub-matrix, but may run the last summation term either for all amino acids or only for some specific groups called regions (R). These regions are often defined in geometric terms and are referred to as core, inner, middle and surface regions. The protein regions (c correspond to core, i to inner, m to middle, and s to surface regions, respectively) are shown in different figures published in previous works [24]. The diameters of the regions, as a percentage of the longest distance $r_{\max}$ with respect to the centre of charge, are 0–25 for region c, 26 to 50 for region i, 51 to 75 for region m, and 76 to 100 for regions. Additionally, we consider the total region (t) that contains all the amino acids in the protein (region diameter 0–100% of $r_{\max}$). Consequently, we can calculate different ξ_m values for a single protein. Each ξ_m values is referred for all the amino acids contained in different Regions R of the protein (c, i, m, s, or t) and all their neighbors placed at topological distance k within this region (k is the named the order). In this work, we calculated in total 5 regions $\times$ 6 (higher order considered) = 30 ξ_m values for each protein.

2.1.1.3. Statistical analysis. The linear mt-QSAR model was constructed using Linear Discriminant Analysis (LDA). All statistical analyses and data exploration were carried out in STATISTICA 6.0 [25]. In so doing, we used the Forward stepwise method implemented in the LDA module of STATISTICA for the selection of variables. In the present work, the independent data test is used by splitting the data randomly in a training series used for a model construction and a cross-validation (CV) one. Let be d_k drugs and ξ_m target molecular descriptors of type kth for different drugs (d), we attempt to develop a simple linear classifier of mt-QSAR type with the general formula:

$$S(DTP)_{pred} = \sum_{k=0}^{ndd} a_k\cdot d_k + \sum_{m=0}^{ntd} b_m\cdot\xi_m + c_0 \quad (2)$$

We used LDA to fit this discriminant function. The model deals with the classification of a compound set with or without affinity on different receptors. A dummy variable Affinity Class (AC) was used as input to codify the affinity. This variable indicates either high (AC = 1) or low (AC = 0) affinity of the drug by the receptor. $S(DPI)_{pred}$ or DPI affinity predicted score is the output of the model and it is a continuous dimensionless score that sorts compounds from low to high affinity to the target coinciding DPIs with higher values of $S(DPI)_{pred}$ and nDPIs with lowest values. In Equation (2), a and b represents the coefficients of the classification function, determined by the LDA module of the STATISTICA 6.0 software package [25]. We used Forward Stepwise algorithm for a variable selection. The statistical significance of the LDA model was determined calculating the p-level (p) of error with Chi-square test. We also inspected the Specificity, Sensitivity, and total Accuracy to determine the quality-of-fit to data in training. The validation of the model was corroborated with external prediction series.

2.1.1.4. ANN analysis. The non-linear mt-QSAR model was constructed using ANN analysis. All models trained were carried out in STATISTICA 6.0 [25]. In so doing, we used a very simple type of ANN called Three Layers Perceptron (MLP-3) to fit this discriminant function. The model deals with the classification of a compound set with or without affinity on different receptors. A dummy variable Affinity Class (AC) was used as input to codify the affinity. This variable indicates either high (AC = 1) or low (AC = 0) affinity of the drug by the receptor. $S(DTP)_{pred}$ or DTP affinity predicted score is the output of the model and it is a continuous dimensionless score that sorts compounds from low to high affinity to the target coinciding DTPs with higher values of $S(DTP)_{pred}$ and nDTPs with lowest values. In Equation (2), b represents the coefficients of the LNN classification function, determined by the ANN module of the STATISTICA 6.0 software package [25]. We used Forward Stepwise algorithm for a variable selection.

In addition, we can explore more complicated non-linear ANNs in order to improve the accuracy of the classifier. We processed our data with different ANNs looking for a better model. Four types of ANNs were used, namely, Probabilistic Neural Network (PNN), Radial Basic Function (RBF), Linear Neural Network (LNN), and Four Layer Perceptron (MLP-4) [26,27]. The quality of all the ANNs (linear or non-linear) was determined calculating values of Specificity, Sensitivity, and total Accuracy to determine the quality-of-fit to data in training. The validation of the model was corroborated with external prediction series. We also reported ROC-curve analysis (ROC curve can be used to select an optimum decision) for both training and validation series [26,28].

2.1.1.5. Data set. The data set was formed by a set of marketed DPIs with known affinity of drugs by targets. This dataset is the same benchmark data used in previous works [2,8–10] in this area and contains all drugs approved by the US FDA. We download this dataset from the public resource called Drug Bank [8,9]{Knox, 2011 #11246; Wishart, 2008 #11249; Wishart, 2006 #11250}. The data set was formed for more than 519 drugs with their respectively 336 targets. Subsequently, we were able to collect above 2337 cases (drug-protein interactions) instead of 519 $\times$ 336 cases. In addition the data set was used to develop ANN models to performance the model. The names or codes for all compounds are depicted in

Table 1SM and Table 2SM of the supplementary material, due to space constraints, as well as the references consulted to compile the data in this table.

2.1.1.6. Complex network construction. We construct a DPIs network in order to achieve the drug and protein affinity with a network approach. Generally in this network, one node may represent a drug or a target. On the other hand, the edges represents the DPIs; express relationships between pairs of drugs with their targets [10]. Anyhow, the nodes representing targets may be of at least two types. In almost all cases reported up to date each target is represented only once in the network. In this class of "static" DPIs network the target is depicted by the node corresponding to the X-ray structure of itself. In this work, we build in total two complex networks. First, we constructed the DPIs networks for the observed data and second, DPIs network predicted by the model. The common steps to construct these networks are:

1. First, using the Excel software in a column we introduce all the proteins, the drugs used quotation marks in our database.
2. Then in another column lists all the cases. At the beginning of this column puts the total number of vertices, there are currently two columns of the name of drug and protein and their corresponding number of vertices.
3. At the end of the columns are placed bows in the first column put the number of vertices for the drug and in another column corresponding to the protein.
4. The file was saved as a .txt format file. After we had renamed the .txt file as a .net file we read it with the CentiBin software [29,30].
5. Using CentiBin we can not only represent the network but also highlight all drugs and targets (nodes) connected by a specific edge or link (DPI). Using this software we can calculate vertex centralities to analyze the relationships between drug targets.

2.2. Illustrative experiments

2.2.1. Synthesis of oxoisoaporphines

2.2.1.1. Synthesis. Synthesis of compounds **1**–**10** has been previously reported by us [31,32]. The 2,3-dihydrooxoisoaporphines 2 and 3 (Scheme 1) were obtained starting from condensation product of 3,4-dimethoxyphenylethylamine (homoveratrylamine) with phthalaldehydic acid. The intermediate compound was subsequently treated with polyphosphoric acid to give the compounds 2 and 3. Treatment with 10% Pd on charcoal over benzene of 2 and 3 yielded the isoaporphines 5 and 6 39. Compound 2 was catalytically hydrogenated over PtO2 at room temperature at 60–70 psi in AcOH affording, in good yield, 9 in which only ring D is saturated 40, 41. By NaBH4 reduction of 2 carbinol 10 is obtained 37. Finally, treatment of 5 with dust zinc and 37% HCl to give the phenolic compound 7. On the other hand, 1 (Scheme 2) was obtained from N-phenethylphthalimide, which was partially reduced with $NaBH_4$ in MeOH and cyclized with hydrochloric acid to give 5,6,8,12b-tetrahydro-8-isoindolo [1,2-a]isoquinolone and this was oxidized with air in the presence of NaOH-MeOH and dimethyl sulfate to afford 1-(2-methoxycarbonylphenyl)-3,4-dihydroisoquinoline, which was directly hydrolyzed with hydrochloric acid to 1-(2-carboxyphenyl)-3,4-dihydroisoquinoline, which, using fuming sulfuric acid, was finally cyclized affording 1. Treatment of 1 with 10% Pd on charcoal over benzene yielded isoaporphine 4 [33]. The Bischler–Napieralski condensation of homoveratrylamine and 2-(benzylbenzoate)chloride afforded a compound characterized as (20-(3,4-dihydro-6,7-dimethoxyisoquinolin-10-yl)phenyl)methylbenzoate that, when it was made to react with an $AcOH/H_2SO_4$ mixture at 100 °C, surprisingly afforded only compound 5-methoxy-6H-dibenzo[de,h] quinolin-6-one compound 8 (Scheme 3) [34,35]; Schemes 1, 2 and 3 are in Fig. 3.

3. Results

3.1. DPIs QSAR predictive models

3.1.1. LDA model

Common physicochemical properties like electrostatic potentials have been demonstrated to be useful on protein QSAR [36,37]. We used these properties as input of our model in addition to drug molecular descriptors. The present is the first mt-QSAR model combining DRAGON and MI to predict the probability with which occur DPIs between a drug and a protein. This type of models lie within the frontiers between classic QSAR for drugs and protein QSAR [38]. Some applications for the present model are the prediction of new drugs, new protein receptors or drug targets, and drug binding sites. Detailed information on the compounds, predicted classification, and probability of affinity on different receptors of the drugs used to seek the model appears in Table 1SM of the supplementary material. Based on the algorithms described in materials and methods the best linear model found was the following:

$$\begin{aligned} S(DTP)_{\text{pred}} = & -0.0004\cdot d_1 - 0.0032\cdot d_2 - 0.0026\cdot d_3 \\ & + 0.0021\cdot d_4 - 0.0019\cdot d_5 - 0.0003\cdot d_6 \\ & - 0.0026\cdot d_7 + 0.1284\cdot d_8 - 0.0253\cdot \xi_1 + 4.2642 \\ & N = 2337 \quad \chi^2 = 2585.53 \quad p-\text{level} < 0.001 \end{aligned} \tag{3}$$

The nomenclature used in the descriptors of the equation is found in Table 1. In this equation, N is the number of cases, χ^2 is the Chi-square and p is the level of error. This model, with 9 variables, classifies correctly 533 out of 588 DPIs (Sensitivity of 77.47%) and 800 out of 810 nDPIs (Specificity of 98.76%). Overall training Accuracy was 88.98%. The validation of the model was carried out by means of external predicting series. The model classifies correctly 252 out of 339 DPIs (74.34%) and 490 out of 510 nDPIs (96.08%) in validation series. Accuracy for validation series (predictability) was 87.37%. These results (Table 2) indicate that we developed an accurate model according to previous reports on the use of LDA in QSAR [39,40].

After observing the above result, we developed another model to incorporate possible interaction effects between drug and protein descriptors, with the aim of improving the model. For this, we used the product between descriptors resulting from the linear mt-QSAR equation. In fact, one interaction effect was entered in the new equation after rerunning forward stepwise analysis. The model equation incorporating interaction effects is as follows:

$$\begin{aligned} S(DTP)_{\text{pred}} = & -0.0044\cdot d_2 + 0.0013\cdot d_9 + 0.0032\cdot d_{10} \\ & + 0.00022\cdot d_6 - 0.0017\cdot d_{11} - 0.0003\cdot d_6 \\ & + 0.0015\cdot d_4\cdot \xi_1 - 1.643 \\ & N = 2337 \quad \chi^2 = 565.4862 \quad p-\text{level} < 0.001 \end{aligned} \tag{4}$$

This model, with 9 variables, classifies correctly 276 out of 339 DPIs (Sensitivity of 81.42%) and 430 out of 510 nDPIs (Specificity of 84.31%). Overall training Accuracy was 83.16%. The validation of the model was carried out by means of external predicting series. The model classifies correctly 539 out of 678 DPIs (79.5%) and 700 out of 810 nDPIs (86.42%) in validation series. Accuracy for validation series (predictability) was 83.27%. These results (Table 2) indicate that we can seek a statistically significant and relatively accurate linear model according to previous reports on the use of LDA in QSAR [39,40]. We can also conclude that even when the stepwise

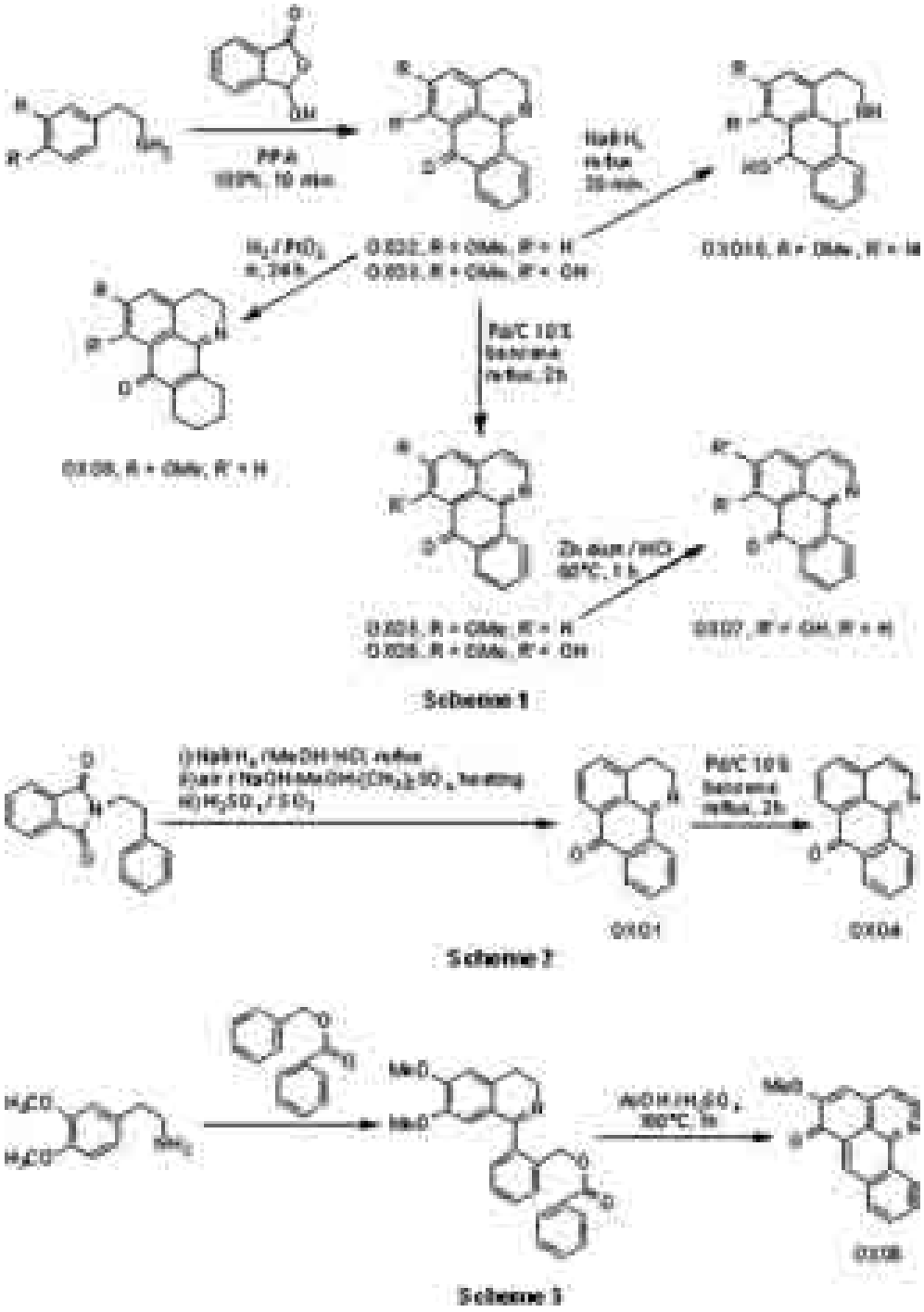


Fig. 3. Oxoisoaporphines derivatives used in this work.

analysis incorporates an interaction effect it does not improve the final model. In conclusion, there is a not-random linear relationship between DPIs and drug + protein descriptors calculated with DRAGON and MI. We can also conclude that we may need to carry out non-linear techniques to improve the model.

3.1.2. Train and validation of 2D MI-DRAGON ANN model

The previous models show good results with a relatively small number of parameters (9 parameters and 7 parameters) and a linear equation for each one. However, as result of the previous section we decided to carry out an ANN analysis to seek a better model using a non-linear method. Four types of ANNs were used, namely, Probabilistic Neural Network (PNN), Radial Basic Function (RBF), Three Layers Perceptron (MLP-3), and Four Layer Perceptron (MLP-4). See, previous works about the use of these ANNs in protein QSAR [2,9]. The Fig. 4 depicts the networks topology for some of the ANN models tested. In general, at least one ANN of every type tested was statically significant. However, one must note that the profiles of each network indicate that many of these are highly non-linear and complicated models.

ANN-QSAR model has been demonstrated before; see, for instance, the works of Fernandez and Caballero [41,42]. We

Table 1
Detailed list of the symbols and description for all parameters present in the model.

Original Descriptor	Descriptor name	Code ID
ATS6v	Broto-Moreau autocorrelation of a topological structure - lag 6/weighted by atomic van der Waals volumes	d_1
GATS2e	Geary autocorrelation - lag 2/weighted by atomic Sanderson electronegativities	d2
BELm1	Lowest eigenvalue n. 1 of Burden matrix/weighted by atomic masses	d3
BELm4	Lowest eigenvalue n. 4 of Burden matrix/weighted by atomic masses	d4
BELm5	Lowest eigenvalue n. 5 of Burden matrix/weighted by atomic masses	d5
GGI1	Topological charge index of order 1	d6
GGI9	Topological charge index of order 9	d7
JGI7*10-3	Mean topological charge index of order7/1000	d8
$^{T}\theta(m)$	Entropy of all aminoacids placed in the middle region and all the neighbors at distance $k \leq 2$	p1
GATS6p	Geary autocorrelation - lag 6/weighted by atomic polarizabilities	d9
BELv5	Lowest eigenvalue n. 5 of Burden matrix/weighted by atomic van der Waals volumes	d10
SEigv	Eigenvalue sum from van der Waals weighted distance matrix	d11
d4p1	(lowest eigenvalue n. 4 of Burden matrix/weighted by atomic masses) (Entropy of all aminoacids placed in the middle region and all the neighbors at distance $k \leq 2$)	d4p1

compare different types of networks to obtain a better model. In Table 2 we show the classification matrix of the different networks. The profiles of networks tested were RBF 1:1-1-1:1 with only one variable; LNN 243:243-1:1, which present many variables, and PNN 243:243-7458-2-2:1, which has a very high number of hidden neurons, see Table 2. After that, the simpler but more accurate ANN model found was an MLP (MLP 21:21-31-1:1) with training Accuracy = 91.06%. This was selected as the best network found because it presents both high accuracy and an adequate number of variables accounting for features relevant for DPIs. This ANN presents 21 inputs variables ($18\ d_k + 3\ \xi_m$). This leads to 21 neurons in first or input layer (I), 31 neurons in the second layer or first hidden layer (H1) and only one neuron (DPI prediction) in the output layer (O). We depict the ROC-curve for MLP 21:21-31-1:1 to show how reliable was the network model developed, see Fig. 5. Notably, almost all the models presented had a ROC-curve higher than 0.5. The model presented an area greater than 0.92. From now on we call the ANN MLP 21:21-31-1:1 as the 2D MI-DRAGON predictor.

3.1.3. 2D MI-DRAGON assembly of CNs for DPIs

A possible application for this model, which is relevant to drug and target screening, is the construction of multi-protein CNs that incorporates protein affinity profile for drugs or the same CNs for DPIs. In order to recall the capacity of 2D MI-DRAGON to predict new CNs of DPIs we selected the same benchmark database used in previous works [2,9]; which includes US FDA approved drugs with their targets. With these goals in mind, we constructed again and manually curated the above-mentioned CN obtaining a graph with 855 vertices or nodes (drugs and proteins) and m = 1016 DPIs (edges). This CN of DPIs have $D = 6.7$; average topological distances D_{ij} between all pairs of nodes. The same as before, we constructed a new CN of DPIs but connecting only pairs of nodes with DPIs predicted by 2D MI-DRAGON. In so doing, we obtained a value of $D = 6.6$ and $m = 907$ DPIs. In Fig. 6 we illustrated visually both CNs (observed and predicted). In first instance, we can see that both networks may have very similar topology (connectivity patterns structure) as measure in terms of D and m.

One way to apply this type of CNs of DPIs for drug screening and drug-target discovery is the calculation of those nodes (drugs or proteins) which are more relevant or important (central) in the graph. For it we can use numerical parameters that quantify the importance of a node in a graph which are called node centralities C_t of type t [43]. The identification of these nodes using node centralities may help us to identify the more relevant drugs or

Table 2
Comparison of LDA and different ANNs classification models.

Model profile	Class	Train			Stat.	Validation		
		%	DPIs	nDPIs	Par.	%	DPIs	nDPIs
2D MI-DRAGON	DPIs	89.38	303	36	Sn	92.18	625	53
MLP	nDPIs	94.12	30	480	Sp	90.12	80	730
21:21-31-1:1	Total	92.23			Ac	91.06		
LDA[a]	DPIs	77.47	155	533	Sn	77.47	155	533
9:9–1:1	nDPIs	98.77	800	10	Sp	98.77	800	10
	Total	88.99			Ac	88.99		
LDA	DPIs	84.31	430	80	Sn	86.42	700	110
6:6–1:1	nDPIs	81.42	63	276	Sp	79.50	139	539
	Total	83.16			Ac	83.27		
PNN	DPIs	63.72	216	123	Sn	67.55	458	220
243:243-7458-2-2:1	nDPIs	74.51	130	380	Sp	67.90	260	550
	Total	70.20			Ac	67.74		
RBF	DPIs	74.04	251	88	Sn	71.24	483	195
1:1-1-1:1	nDPIs	72.55	140	370	Sp	80.25	160	650
	Total	73.14			Ac	76.14		
LNN	DPIs	84.37	286	53	Sn	80.09	543	135
243:243-1:1	nDPIs	100.00	0	510	Sp	81.48	150	660
	Total	93.76			Ac	80.85		

DPIs: Drug-Target Pairs for compounds with high affinity; nDPIs: Drug-Target Pair for compounds with non-affinity; Stat. is statistics, Par. is parameter.

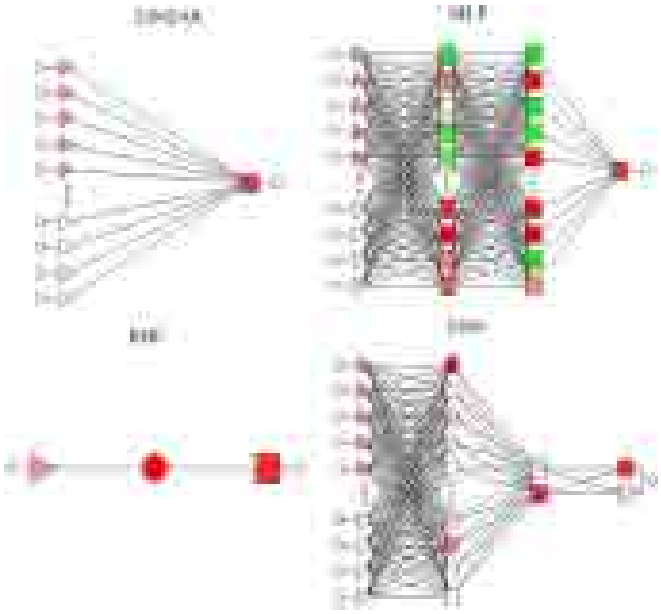


Fig. 4. Generic Topology of ANN models trained in this work.

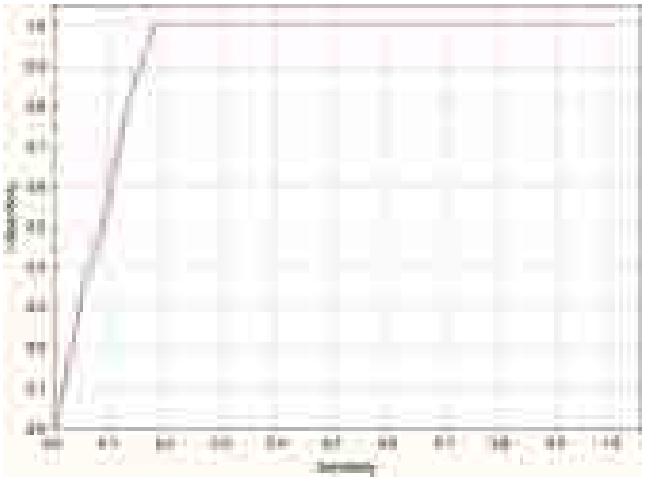


Fig. 5. ROC Curve for 2D MI-DRAGON predictor (red = train series, blue = validation series). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

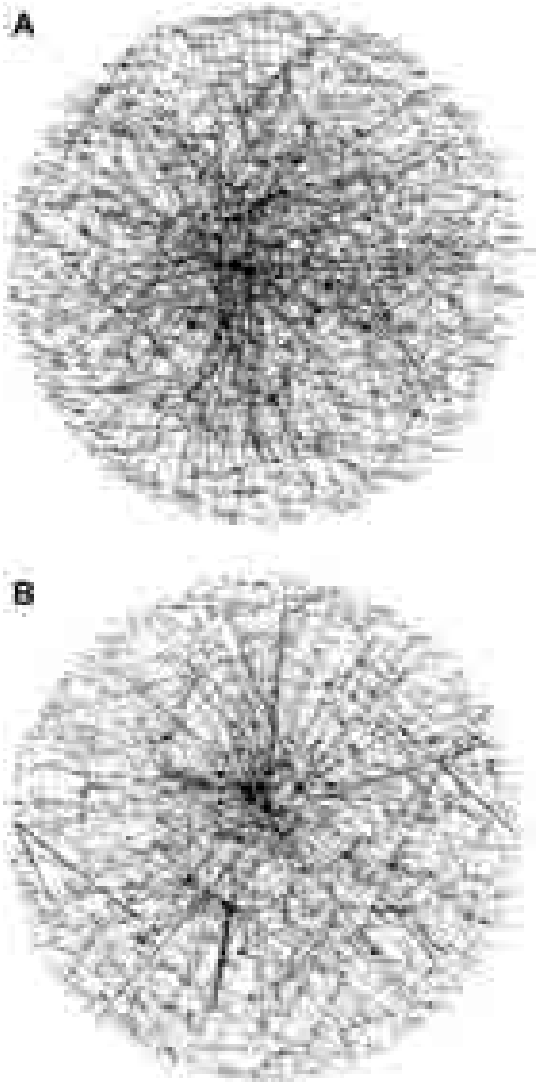


Fig. 6. Observed vs. Predicted drug-target complex networks.

proteins in analogy to similar procedures developed for PINs; networks of Protein–Protein Interactions (PPIs) [44]. In Table 3 we show the predicted results of both node degree centrality (C_δ) and closeness centrality (C_{clo}) for proteins and drugs present in the database. The parameter C_δ measures the local importance of a node by counting the number of nodes directly attached to him [44]. Conversely, C_{clo} measures the global importance of a node in a CN by taking in consideration the inverse of the sum of D_{ij} ($C_{clo} = 1/\Sigma D_{ij}$) [45]. Consequently, the higher C_δ the higher is the local importance of the node but the higher C_{clo} the lower is the global importance of the node. For instance, the protein 1HA2 with lower $C_{clo} = 0.0004$ and a $C_\delta = 47$ is the more important protein both locally and globally in this CN. This means that this protein is both locally and globally important because it is the target of many drugs (high C_δ). This is a very interesting result; because 1HA2 is the protein Human Serum Albumin (HSA). HSA is an abundant transport protein found in plasma that binds a wide variety of drugs in two primary binding sites (I and II) and can have a significant impact on their pharmacokinetics and is characterized by its surprising capacity to bind a large variety of biologically active molecules [46,47]. Depending on our aims the more important nodes in pharmacological terms not necessarily have to be the more central in the graph (those with higher C_δ and lower C_{clo}), see Table 3.

3.2. Theoretic-experimental study using 2D MI-DRAGON predictor

Finally, we illustrated in one theoretic-experimental study the practical use of 2D MI-DRAGON. We reported the prediction, synthesis, and pharmacological assay of 10 different oxoisoaporphines with MAO-A inhibitory activity.

3.2.1. Pharmacological assay for oxoisoaporphines

The test drugs (new compounds and reference inhibitors) they were unable to react directly with the Amplex® Red reagent, excluding any interference in the measurements. In our experiments and under our experimental conditions, hMAO-A displayed a Michaelis constant (K_m) equal to 457.17 ± 38.62 μM and a maximum reaction velocity (V_{max}) equal to 185.67 ± 12.06 nmol/min/mg protein whereas hMAO-B showed a K_m equal to 220.33 ± 32.80 μM and a V_{max} equal to 24.32 ± 1.97 nmol/min/mg protein ($n = 5$). Most tested drugs concentration-dependently and selectively inhibited the enzymatic control activity of MAO-A (see Table 4).

3.2.2. Pharmacological studies: determination of hMAO isoform activity

The potential effects of the test drugs on hMAO activity were investigated by measuring their effects on the production of hydrogen peroxide from *p*-tyramine (a common substrate for both hMAO-A and hMAO-B), using the Amplex® Red MAO assay kit (Molecular Probes, Inc., Eugene, Oregon, USA) and microsomal MAO isoforms prepared from insect cells (BTI-TN-5B1-4) infected with recombinant baculovirus containing cDNA inserts for hMAO-A or hMAO-B (Sigma–Aldrich Química S.A., Alcobendas, Spain). The production of H_2O_2 catalysed by MAO isoforms can be detected using 10-acetyl-3,7-dihydroxyphenoxazine (Amplex® Red reagent), a non-fluorescent and highly sensitive probe that reacts with H_2O_2 in the presence of horseradish peroxidase to produce a fluorescent product: resorufin. In this study hMAO activity was evaluated using the above-mentioned fluorimetric method following the general procedure previously described by us [48]. Briefly, 0.1 ml of sodium phosphate buffer (0.05 M, pH 7.4) containing various concentrations of the test drugs (new compounds or reference inhibitors) and adequate amounts of recombinant hMAO-A or hMAO-B required

Table 3
Results of node degree and closeness centrality for top-20 proteins and drugs.

Rank	PDB	Function	C_δ	C_{clo}	Drugs	Activity	C_δ	C_{clo}
1	1HA2	Human serum albumin	44	4.798	NADH	Energy production	35	3.83
2	1BNA	B-DNA dodecamer	36	4.153	Simvastatin	hypolipidemic drug	18	4.219
3	1R5K	Estrogen receptor	27	3.774	Dimethyl sulfoxide	Solvent	15	4.146
4	1EMI	Ribosomal protein s8	16	2.989	Atorvastatin	hypolipidemic drug	11	4.085
5	1CZM	Carbonic anhydrase I	14	2.573	Pyridoxal Phosphate	transamination reactions	9	2.994
6	1MO8	ATPase	14	3.383	Adenosine monophosphate	Energy production	8	3.484
7	1NHZ	Glucocorticoid receptor	14	2.776	Minocycline	Antibiotic	8	3.477
8	1SQN	Progesterone receptor	13	2.803	Menadione	Protrombin synthesis	7	3.266
9	1T9N	Carbonic anhydrase II	13	2.584	Vitamin A	Antioxidant	7	2.789
10	1UZF	Angiotensin enzyme	13	3.327	Alitretinoin	Antineoplastic agent	6	2.895
11	1BYW	HERG potassium channel N	11	3.195	Halothane	Anesthesic	6	3.946
12	1E3G	Androgen receptor	11	3.621	L-Proline	support tissues of the body	6	2.919
13	1VRU	HIV-1 reverse transcriptase	11	2.929	Acitretin	Psoriasis treatment	5	3.897
14	1ODW	HIV-1 protease	10	3.439	Etretinate	Psoriasis treatment	5	2.894
15	1ZNC	Carbonic Anhidrase IV	10	2.58	Acetozolamide	Carbonic Anhidrase inhibitor	5	2.45
16	1F5F	Human sex hormone-binding globulin	9	3.687	Sulindac	non-steroidal anti-inflammatory drug	5	3.885
17	1HWL	HMG-CoA reductase	9	3.731	Adapalene	anti-inflammatory drug	4	2.892
18	1A8M	Tumor necrosis factor alpha	8	4.085	Cyclothiazide	antihypertensive	4	2.907
19	1TB5	Phosphodiesterases	8	3.209	Dromostanolone	Antineoplasic Agent	4	3.623
20	1TQN	Cytochrome P450 3A4	8	3.592	Estradiol	Hormone	4	3.24

and adjusted to obtain in our experimental conditions the same reaction velocity, i.e., to oxidize (in the control group) 165 pmol of *p*-tyramine/min (hMAO-A: 1.1 μg protein; specific activity: 150 nmol of *p*-tyramine oxidized to *p*-hydroxyphenylacetaldehyde/min/mg protein; hMAO-B: 7.5 μg protein; specific activity: 22 nmol of *p*-tyramine transformed/min/mg protein) were incubated for 15 min at 37 °C in a flat-black-bottom 96-well microtiter (microtest™ plate, BD Biosciences, Franklin Lakes, NJ, USA) placed in the dark multi-mode microplate reader chamber. After this incubation period, the reaction was started by adding (final concentrations) 200 μM Amplex Red reagent, 1 U/ml horseradish peroxidase and 1 mM *p*-tyramine. The production of H_2O_2 and, consequently, of resorufin was quantified at 37 °C in a multi-mode microplate reader (Fluostar Optima, BMG Labtech GmbH, Offenburg,Germany), based on the fluorescence generated (excitation, 545 nm, emission, 590 nm) over a 15 min period, in which the fluorescence increased linearly.

Control experiments were carried out simultaneously by replacing the test drugs (new compounds and reference inhibitors) with appropriate dilutions of the vehicles. In addition, the possible capacity of the above test drugs to modify the fluorescence generated in the reaction mixture due to non-enzymatic inhibition (e.g., for directly reacting with Amplex® Red reagent) was determined by adding these drugs to solutions containing only the Amplex® Red reagent in a sodium phosphate buffer. To determine the kinetic parameters of hMAO-A and hMAO-B (K_m and V_{max}), the

Table 4
Inhibitory activity of different isooxoaporphine derivatives (**OXO 1- OXO 10**).

Compounds	hMAO-A (IC_{50} μM)	hMAO-B (IC_{50} μM)	SI
OXO 1	27.32 ± 1.18[a]	>100	>3.7[b]
OXO 2	0.014 ± 0.00034[a]	>100	>7,143[b]
OXO 3	0.044 ± 0.0022[a]	>100	>2,273[b]
OXO 4	0.72 ± 0.05[a]	>100	>139[b]
OXO 5	0.00083 ± 0.000049[a]	>100	>120,482[b]
OXO 6	1.23 ± 0.08[a]	>100	>81[b]
OXO 7	0.035 ± 0.0013[a]	>100	>2,857[b]
OXO 8	0.013 ± 0.00054[a]	>100	>7,692[b]
OXO 9	2.12 ± 0.07[a]	>50	>24[b]
OXO 10	9.81 ± 0.26[a]	>50	>5.1[b]
Clorgyline	0.0040 ± 0.00025[a]	63.41 ± 1.20	15,852
Deprenyl	68.73 ± 4.21[a]	0.017 ± 1.96	0.00025
Iproniazid	6.56 ± 0.76[a]	7.54 ± 0.36	1.15
Moclobemide	361.38 ± 19.37[a]	*	>2.8[b]

SI: hMAO-A selectivity index = $IC_{50\ (hMAO-B)}/IC_{50\ (hMAO-A)}$. Each IC_{50} value is the mean ± S.E.M. from five experiments. * Inactive at 1 mM (highest concentration tested). Level of statistical significance.

[a] $P < 0.01$ versus the corresponding IC_{50} values obtained against hMAO-B, as determined by ANOVA/Dunnett's.

[b] Values obtained under the assumption that the corresponding IC_{50} against MAO-B is the highest concentration tested (10 μM or 100 μM).

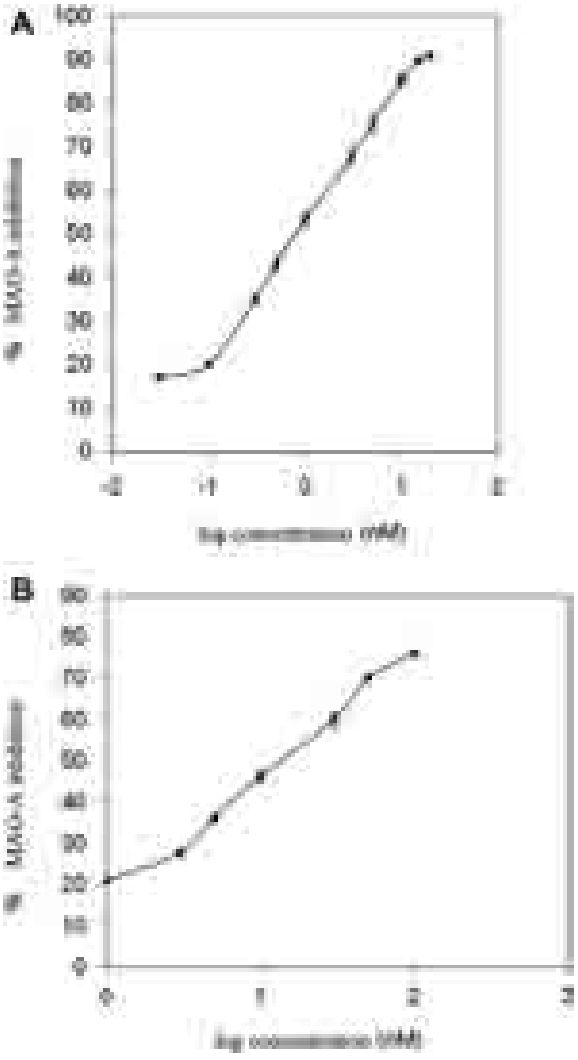


Fig. 7. Concentration-response curves for Oxoisoaporphines (OXO 5 is A, OXO 8 is B).

Table 5
Prediction of oxoisoaporphines analogues with 2D MI-DRAGON predictor.

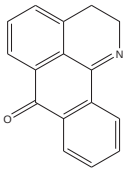
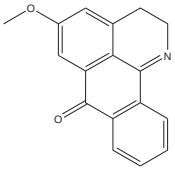
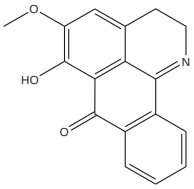
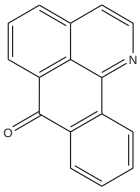
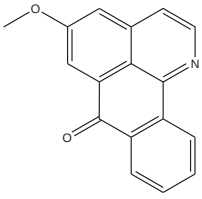
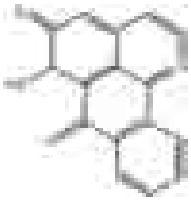
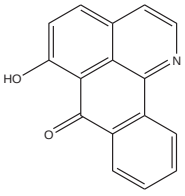
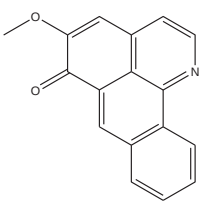
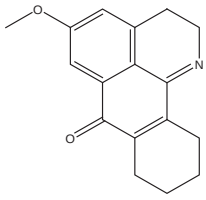
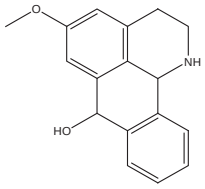
Compound			MAO-A	
Drug	OC	PC	$S(DPI)_{pred}$	Structure
OXO 1	0	1	0.981	
OXO 2	1	1	0.996	
OXO 3	1	1	0.997	
OXO 4	1	1	0.981	
OXO 5	1	1	0.996	
OXO 6	1	1	0.997	
OXO 7	1	1	0.992	
OXO 8	1	1	0.997	

Table 5 (*continued*)

Compound			MAO-A	
OXO 9	1	1	0.996	
OXO 10	1	1	0.996	

OC = observed class and PC = Predicted class, OC = 1 if compound $IC_{50} < 10$ μM and PC = 1 if the DPI probability predicted for pair drug-MAO-a enzyme p(MAO-a) > 0.5 (2Z5X is PDB ID of MAO-A used to predict *p*-values).

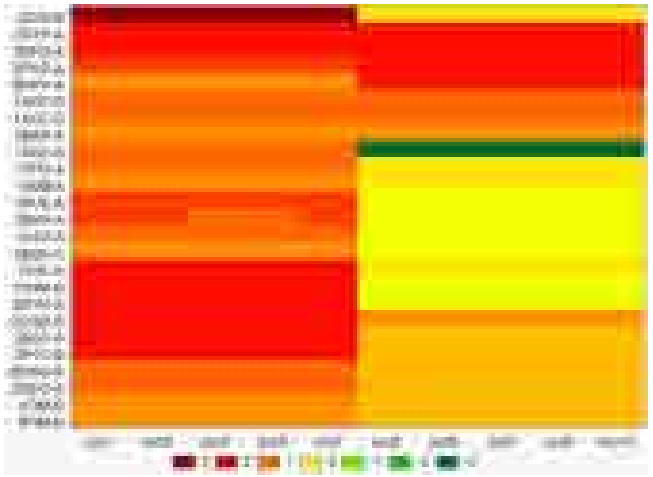
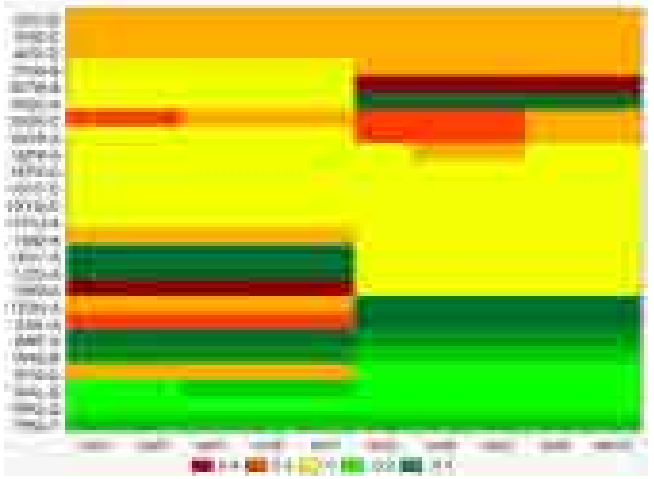
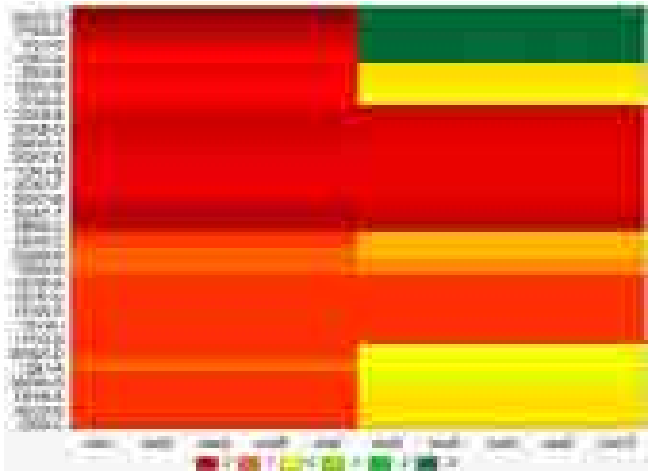
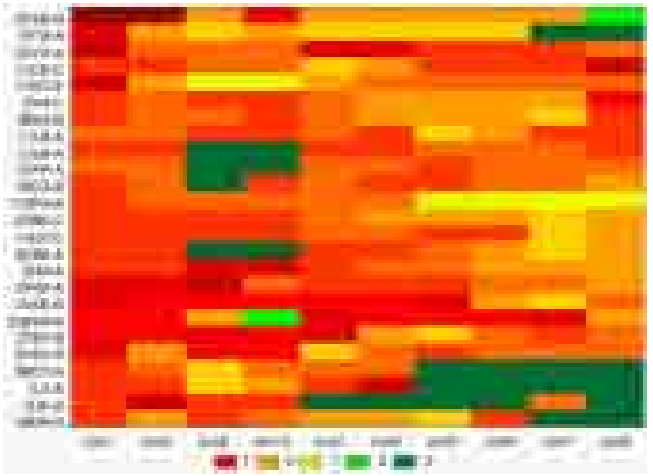
corresponding enzymatic activity of both isoforms was evaluated (under the experimental conditions described above) in presence of a number (a wide range) of *p*-tyramine concentrations. The specific fluorescence emission (used to obtain the final results) was calculated after subtraction of the background activity, which was determined from vials containing all components except the MAO isoforms, which were replaced by a sodium phosphate buffer solution.

The results show that compounds called OXO 5 and OXO 8 have IC_{50} values in the range of pM and nM respectively, much lower than the reference inhibitor of MAO-A, which can be seen in the concentration-response curves for each compound (see Fig. 7). Fig. 7A shows the curve "concentration-response representative of the inhibitory effects produced by the compound OXO 5 (IC_{50} = 835.75 ± 49.75 pM) on the enzymatic activity of human recombinant MAO-A. Fig. 7B shows the curve "concentration-response representative of the inhibitory effects produced by the compound OXO 8 (IC_{50} = 13.36 ± 0.54 nM) on the enzymatic activity of human recombinant MAO-A. Each point in both represents the mean ± SEM (indicated by vertical lines) of at least five experiments. Other compounds showed IC_{50} values in the range of nM to μM, and were very selective inhibitors of MAO-A isoform.

3.2.3. 2D MI-DRAGON prediction of oxoisoaporphines vs. MAO-A

In this *in silico* experiment we used 2D MI-DRAGON to predict the interaction of the new oxoisoaporphines with MAO-A. For it, we downloaded the 3D structure of MAO-A protein with PDB ID 2Z5X and calculated their structural parameters with MI. We also generated the SMILE codes for these compounds and calculated their structural parameters with DRAGON. After that, we predicted their propensity to undergo DPIs with MAO-A using as inputs for the 2D MI-DRAGON predictor the structural parameters of both the drugs and the protein. In Table 5 we confront the results obtained using this model and the outcomes of the pharmacological assay. The compound Clorgyline a known selective inhibitor for MAO-A was used as control. We consider the observed class for active compounds OC = 1 if compound $IC_{50} < 10$ μM this cut-off is in the similar range than other used in previous works [49,50]. As we can see in this table all the compounds oxoisoaporphines analogs present some activity, but we found an interesting result in the pharmacological assays. The most active compound in the pharmacological assay (OC = 1) was compound OXO 5 predicted as

Table 6
TWJ analysis for oxoisoaporphines derivatives *vs.* parasite targets.

Trypanosome	Val.	P	Val.	Leishmania
	10	V	10	
	50	C	50	
	0.41	Tc	0.116	
	13	B	18	
	0.68	M	−0.11	
	0.83	Sd	0.23	
Plasmodium	**Val.**	**P**	**Val.**	**Toxoplasma**
	10	V	10	
	60	C	50	
	0.68	Tc	0.58	
	10	B	44	
	0.90	M	−0.15	
	1.36	Sd	1.17	

Number of variables (V = 10 oxoisoaporphine compounds); Number of cases: (*C* = 50–60 proteins) Threshold computed from data (Tc); Number of blocks (B); Total Sample Mean (M); Standard Deviation (Sd); colors in the figure are only illustrative, red score means a higher value and green value a lower score, but in all cases are nDPI score values.

active with PC = 1 and high score $S(DPI)_{pred} = 0.99$. The OXO 8 compound that are active in pharmacological assays (OC = 1) were also predicted as active against MAO A (PC = 1) and high score $S(DPI)_{pred} = 0.99$.

3.2.4. 2D MI-DRAGON screening of oxoisoaporphines vs. US FDA proteins

An additional use of 2D MI-DRAGON was to carry out the "*in silico*" or virtual screening of the new compounds with respect to all other targets previously approved by US FDA [51]. It may help to found new targets for these drugs or discard possible toxicological effects depending on the other targets predicted and/or discarded for these compounds. This type of experiment is of the major importance due to the cost in terms of animal sacrifice, time, materials and human resources of the experimental assay of all compounds against all these targets, see recent reviews by Duardo-Sanchez et al. [52–55]. In fact, over a decade, the US FDA has been engaged in the applied research, development, and evaluation of computational toxicology methods used to support the safety evaluation of a diverse set of regulated products. The basis for evaluating computational toxicology methods is multi-factorial, including the potential for increased efficiency, reduction in the numbers of animals used, lower costs, and the need to explore emerging technologies that support the goals of the US FDA's Critical Path Initiative (e.g. to make decision support information available early in the drug review process) [56].

In this experiment, we downloaded the 3D structure of all proteins that are targets of US FDA approved drugs. Next, we calculated the structural parameters of all these proteins with MI. We also generated the SMILE codes for these compounds and calculated their structural parameters with DRAGON. After that, we predicted their propensity to undergo DPIs with all US FDA proteins using as inputs for the 2D MI-DRAGON predictor the structural parameters of both the drugs and proteins. We depict in Table 3SM, all proteins in FDA dataset predicted *vs.* the 10 oxoisoaporphine derivatives. We found that overall the 10 derivatives were predicted as non-active (low DPIs scores) against almost all proteins in the FDA database. Consequently, 2D MI-DRAGON predicts a high the selectivity of the new oxoisoaporphines derivatives as MAO A inhibitors. We can reach this goal because the model predicts these compounds as non-active with respect to all proteins that are targets of FDA drugs.

3.2.5. 2D MI-DRAGON study of oxoisoaporphines selectivity for parasite proteins

An additional use of 2D MI-DRAGON was to predict the selectivity of the new oxoisoaporphine derivatives with respect to other targets in different organisms. In a previous work we reported antiplasmodial activity ($IC_{50} = 1.45$ μM) for compound **8**, whereas **5** exhibited a lower activity ($IC_{50} \sim 10$ μM), **2**, **4** and **6** had a moderate effect ($IC_{50} \sim 50$ μM) and **1, 3** and **7** possessed a very low activity ($IC_{50} > 80$ μM) against the same parasite [9]. In this previous work, we also predicted the possible activity of these compounds against a collection of proteins present in *Plasmodium falciparum*. All predictions were carrying out using the web server NL MIND-BEST. The outputs of these studies were used as inputs for a Two-Way-joining Analysis (TWJ) of these results. These results obtained in the previous work indicated a low average tendency of these drugs to bind the selected *P. falciparum* proteins. However, NL MIND-BEST predicted that compounds **4**, **5**, and **8** have higher propensities to interact with some *P. falciparum* proteins. This result was significant because the compound **8** has shown activity against *P. falciparum* in the experimental assay carry out in the previous work. This previous study paved the way to new predictive studies of another potential targets for this class of compounds in *P. falciparum* and/or other parasite species.

In this sense, we decided to carry out here a predictive study of potential targets for oxoisoaporphines using 2D MI-DRAGON predictor. For it, first we first we downloaded the 3D structure of a large set of 1660 human parasite protein chains from PDB. In total, we studied 627 protein chains of Trypanosome, 230 of Leishmania, 136 of Toxoplasma, as well as other 626 of Plasmodium (not studied in the previous study). Next, we calculated the structural parameters of all these proteins with MI. We also generated the SMILE codes for these compounds and calculated their structural parameters with DRAGON, the same than in previous section. After that, we predicted the propensity of oxoisoaporphines to undergo DPIs with all selected 1660 human parasite proteins. Next we developed a Two-Way-Joining (TWJ) cluster analysis in order to facilitate information extraction and graphical depiction from raw tables of drug *vs.* target results. In Table 6 we depict four figures with TWJ hot maps. Each map displays the standardized DPI score values for the 10 oxoisoaporphine derivatives against above 50-60 proteins of one different parasite selected out of the total proteins studied. We found that overall the 10 derivatives were predicted with low DPIs scores against almost all proteins studied (colors in the figure are only illustrative, red score means a higher value and green value a lower score, but in all cases are nDPI score values). This result coincides with the predictions carry out using NL MIND-BEST web server for *P. falciparum*. It points to a high selectivity of these drugs for one specific target in *P. falciparum* (see previous work) with low probability of action against other targets in *P. falciparum* or the other human parasites studied.

4. Conclusions

It is possible to seek excellent predictors for DPIs using as input structural parameters of drugs and proteins calculated with different programs and combined with ANN models. The 2D MI-DRAGON predictor based on structural parameters of drugs calculated with DRAGON and parameters of proteins calculated with MI correctly predicts PDIs of 500 + different drugs approved by US FDA with Accuracy >90%. 2D MI-DRAGON predictor is also useful to assemble CNs of DPIs. These CNs computationally assemble offer an alternative to discover new drugs or targets, and explore the selectivity and toxicity of drugs. In this work, we exemplified these conclusions through the experimental-theoretical study of the MAO A activity of new oxoisoaporphines.

Acknowledgments

González-Díaz H. and Sobarzo-Sanchez E. thank sponsorships for a research position at the University of Santiago de Compostela from the Isidro Parga Pondal Program, Xunta de Galicia. Prado–Prado F. thanks sponsorships for research position at the University of Santiago de Compostela from Angeles Alvariño, Xunta de Galicia.

Appendix. Supplementary data

Supplementary data related to this article can be found online at doi:10.1016/j.ejmech.2011.09.045.

References

[1] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, M. Kanehisa, Prediction of drug-target interaction networks from the integration of chemical and genomic spaces, Bioinformatics 24 (2008) i232–i240.
[2] F. Prado-Prado, X. Garcia-Mera, P. Abeijon, N. Alonso, O. Caamano, M. Yanez, T. Garate, M. Mezo, M. Gonzalez-Warleta, L. Muino, F.M. Ubeira, H. Gonzalez-Diaz, Using entropy of drug and protein graphs to predict FDA drug-target network: theoretic-experimental study of MAO inhibitors and hemoglobin peptides from Fasciola hepatica, Eur. J. Med. Chem. 46 (2011) 1074–1094.

[3] A.M. Helguera, R.D. Combes, M.P. Gonzalez, M.N. Cordeiro, Applications of 2D descriptors in drug design: a DRAGON tale, Curr. Top. Med. Chem. 8 (2008) 1628–1655.
[4] E. Estrada, E. Molina, D. Nodarse, E. Uriarte, Structural contributions of substrates to their binding to P-Glycoprotein. A TOPS-MODE approach, Curr. Pharm. Des 16 (2010) 2676–2709.
[5] Y. Marrero-Ponce, G.M. Casanola-Martin, M.T. Khan, F. Torrens, A. Rescigno, C. Abad, Ligand-based computer-aided discovery of tyrosinase inhibitors. Applications of the TOMOCOMD-CARDD method to the elucidation of new compounds, Curr. Pharm. Des 16 (2010) 2601–2624.
[6] H. Gonzalez-Diaz, A. Duardo-Sanchez, F.M. Ubeira, F. Prado-Prado, L.G. Perez-Montoto, R. Concu, G. Podda, B. Shen, Review of MARCH-INSIDE & complex networks prediction of drugs: ADMET, anti-parasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers, Curr. Drug Metab. 11 (2010) 379–406.
[7] F.J. Prado-Prado, F. Borges, L.G. Perez-Montoto, H. Gonzalez-Diaz, Multi-target spectral moment: QSAR for antifungal drugs vs. different fungi species, Eur. J. Med. Chem. 44 (2009) 4051–4056.
[8] H. Gonzalez-Diaz, F. Prado-Prado, X. Garcia-Mera, N. Alonso, P. Abeijon, O. Caamano, M. Yanez, C.R. Munteanu, A. Pazos, M.A. Dea-Ayuela, M.T. Gomez-Munoz, M.M. Garijo, J. Sansano, F.M. Ubeira, MIND-BEST: web server for drugs and target discovery; design, synthesis, and assay of MAO-B inhibitors and theoretical-experimental study of G3PDH protein from Trichomonas gallinae, J. Proteome Res. 10 (2011) 1698–1718.
[9] H. Gonzalez-Diaz, F. Prado-Prado, E. Sobarzo-Sanchez, M. Haddad, S. Maurel Chevalley, A. Valentin, J. Quetin-Leclercq, M.A. Dea-Ayuela, M. Teresa Gomez-Munos, C.R. Munteanu, J. Jose Torres-Labandeira, X. Garcia-Mera, R.A. Tapia, F.M. Ubeira, NL MIND-BEST: a web server for ligands and proteins discovery–theoretic-experimental study of proteins of Giardia lamblia and new compounds active against Plasmodium falciparum, J. Theor. Biol. 276 (2011) 229–249.
[10] M.A. Yildirim, K.I. Goh, M.E. Cusick, A.L. Barabasi, M. Vidal, Drug-target network, Nat. Biotechnol. 25 (2007) 1119–1126.
[11] Talete srl, in: DRAGON for Windows (Software for Molecular Descriptor Calculations) (2005).
[12] J. Kirchmair, P. Markt, S. Distinto, D. Schuster, G.M. Spitzer, K.R. Liedl, T. Langer, G. Wolber, The Protein Data Bank (PDB), its related services and software tools as key components for in silico guided drug discovery, J. Med. Chem. 51 (2008) 7021–7040.
[13] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A.C. Guo, D.S. Wishart, DrugBank 3.0: a comprehensive resource for 'omics' research on drugs, Nucleic Acids Res. 39 (2011) D1035–D1041.
[14] D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, DrugBank: a knowledgebase for drugs, drug actions and drug targets, Nucleic Acids Res. 36 (2008) D901–D906.
[15] D.S. Wishart, C. Knox, A.C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, DrugBank: a comprehensive resource for in silico drug discovery and exploration, Nucleic Acids Res. 34 (2006) D668–D672.
[16] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors (2000).
[17] E. Papa, F. Villa, P. Gramatica, Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in Pimephales promelas (fathead minnow), J. Chem. Inf. Model. 45 (2005) 1256–1266.
[18] F.P.-P.a.F.M.U. Humberto González-Díaz, Predicting Antimicrobial drugs and targets with the MARCH INSIDE approach, Curr. Top. Med. Chem. 8 (2008) 1676–1690.
[19] R. Concu, G. Podda, E. Uriarte, H. Gonzalez-Diaz, Computational chemistry study of 3D-structure-function relationships for enzymes based on Markov models for protein electrostatic, HINT, and van der Waals potentials, J. Comput. Chem. 30 (2009) 1510–1520.
[20] H. González-Díaz, Y. González-Díaz, L. Santana, F.M. Ubeira, E. Uriarte, Proteomics, networks and connectivity indices, Proteomics 8 (2008) 750–778.
[21] M. Vassura, L. Margara, P. Di Lena, F. Medri, P. Fariselli, R. Casadio, Reconstruction of 3D structures from protein contact maps, IEEE/ACM Trans. Comput. Biol. Bioinform. 5 (2008) 357–367.
[22] Y. Rodriguez-Soca, C.R. Munteanu, J. Dorado, J. Rabuñal, A. Pazos, González-Díaz, Plasmod-PPI: a web-server predicting complex biopolymer targets in plasmodium with entropy measures of protein-protein interactions, Polymer 50 (2009). doi:10.1016/j.polymer.2009.1011.1029.
[23] Y. Rodriguez-Soca, C.R. Munteanu, F.J. Prado-Prado, J. Dorado, A. Pazos Sierra, H. Gonzalez-Diaz, Trypano-PPI: a web server for prediction of unique targets in trypanosome proteome by using electrostatic parameters of protein-protein interactions, J. Proteome Res. (2009). doi:10.1021/pr900827b.
[24] H. González-Díaz, Y. Pérez-Castillo, G. Podda, E. Uriarte, Computational chemistry Comparison of Stable/Nonstable protein Mutants classification models based on 3D and topological indices, J. Computational Chem. 28 (2007) 1990–1995.
[25] StatSoft.Inc, STATISTICA (Data Analysis Software System), version 6.0. Statsoft, Inc, 2002.www.statsoft.com.
[26] F.J. Prado-Prado, X. Garcia-Mera, H. Gonzalez-Diaz, Multi-target spectral moment QSAR versus ANN for antiparasitic drugs against different parasite species, Bioorg. Med. Chem. 18 (2010) 2225–2231.
[27] Y. Rodriguez-Soca, C.R. Munteanu, J. Dorado, A. Pazos, F.J. Prado-Prado, H. Gonzalez-Diaz, Trypano-PPI: a web server for prediction of unique targets in trypanosome proteome by using electrostatic parameters of protein-protein interactions, J. Proteome Research 9 (2010) 1182–1190.
[28] F.J. Prado-Prado, H. Gonzalez-Diaz, L. Santana, E. Uriarte, Unified QSAR approach to antimicrobials. Part 2: predicting activity against more than 90 different species in order to halt antibacterial resistance, Bioorg. Med. Chem. 15 (2007) 897–902.
[29] B.H. Junker, D. Koschutzki, F. Schreiber, Exploration of biological network centralities with CentiBiN, BMC Bioinform. 7 (2006) 219.
[30] D. Koschützki pp. CentiBiN Version 1.4.2, in: CentiBiN Version 1.4.2, Centralities in Biological Networks © 2004-2006 Dirk Koschützki Research Group Network Analysis, IPK Gatersleben, Germany, 2006.
[31] G.N. Walker, D. Alkalay, New synthesis of 4-aryl-2,3-dihydro- and 2,3,4,5-tetrahydro-2(1H)-benzazepines and corresponding 1,3-diones, J. Org. Chem. 36 (1971) 461–465.
[32] E. Sobarzo-Sánchez, B.K. Cassels, C. Jullian, L. Castedo, Complete 1H and 13C NMR spectral assignment of hydrogenated oxoisoaporphine derivatives, Mag. Reson. Chem. 41 (2003) 545–548.
[33] E. Sobarzo-Sanchez, J. De la Fuente, L. Castedo, Synthesis and total assignment of 1H and 13C NMR spectra of new oxoisoaporphines by long-range hetero-nuclear correlations, Mag. Reson. Chem. 43 (2005) 1080–1083.
[34] E. Sobarzo-Sánchez, B.K. Cassels, L. Castedo, Complete structural and spectral assignment of oxoisoaporphines by HMQC and HMBC experiments, Mag. Reson. Chem. 41 (2003) 296–300.
[35] E. Sobarzo-Sánchez, B.K. Cassels, C. Jullian, L. Castedo, An Expedient synthesis of Unusual oxoisoaporphine and Annelated Quinoline derivatives, Synlett. (2005).
[36] O. Ivanciuc, N. Oezguen, V.S. Mathura, C.H. Schein, Y. Xu, W. Braun, Using property based sequence motifs and 3D modeling to determine structure and functional regions of proteins, Curr. Medicinal Chemistry 11 (2004) 583–593.
[37] C.H. Schein, O. Ivanciuc, W. Braun, Common physical-chemical properties correlate with similar structure of the IgE epitopes of peanut allergens, J. Agri. Food Chem. 53 (2005) 8752–8759.
[38] H. Gonzalez-Diaz, L. Saiz-Urra, R. Molina, L. Santana, E. Uriarte, A model for the recognition of protein kinases based on the entropy of 3D van der Waals interactions, J. Proteome Res. 6 (2007) 904–908.
[39] Y.M. Alvarez-Ginarte, Y. Marrero-Ponce, J.A. Ruiz-Garcia, L.A. Montero-Cabrera, J.M. Vega, P. Noheda Marin, R. Crespo-Otero, F.T. Zaragoza, R. Garcia-Domenech, Applying pattern recognition methods plus quantum and physicochemical molecular descriptors to analyze the anabolic activity of structurally diverse steroids, J. Computational Chem. (2007).
[40] A.H. Morales, J.E. Rodríguez-Borges, X. García-Mera, F. Fernández, M.N. Dias-Sueiro-Cordeiro, Probing the Anticancer activity of Nucleoside analogues: a QSAR model approach using an Internally Consistent training set, J. Med. Chem. 50 (2007) 1537–1545.
[41] M. Fernandez, J. Caballero, A. Tundidor-Camba, Linear and nonlinear QSAR study of N-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivatives as matrix metalloproteinase inhibitors, Bioorg. Med. Chem. 14 (2006) 4137–4150.
[42] J. Caballero, M. Fernandez, Linear and nonlinear modeling of antifungal activity of some heterocyclic ring derivatives using multiple linear regression and Bayesian-regularized neural networks, J. Mol. Model. 12 (2006) 168–181.
[43] H. Gonzalez-Diaz, Y. Gonzalez-Diaz, L. Santana, F.M. Ubeira, E. Uriarte, Proteomics, networks and connectivity indices, Proteomics 8 (2008) 750–778.
[44] H. Jeong, S.P. Mason, A.L. Barabasi, Z.N. Oltvai, Lethality and centrality in protein networks, Nature 411 (2001) 41–42.
[45] E. Estrada, Virtual identification of essential proteins within the protein interaction network of yeast, Proteomics 6 (2006) 35–40.
[46] O. Deeb, M.C. Rosales-Hernandez, C. Gomez-Castro, R. Garduno-Juarez, J. Correa-Basurto, Exploration of human serum albumin binding sites by docking and molecular dynamics flexible ligand-protein interactions, Biopolymers 93 (2010) 161–170.
[47] I. Petitpas, A.A. Bhattacharya, S. Twine, M. East, S. Curry, Crystal structure analysis of warfarin binding to human serum albumin: anatomy of drug site I, J. Biol. Chem. 276 (2001) 22804–22809.
[48] F. Chimenti, E. Maccioni, D. Secci, A. Bolasco, P. Chimenti, A. Granese, S. Carradori, S. Alcaro, F. Ortuso, M. Yanez, F. Orallo, R. Cirilli, R. Ferretti, F. La Torre, Synthesis, stereochemical identification, and selective inhibitory activity against human monoamine oxidase-B of 2-methylcyclohexylidene-(4-arylthiazol-2-yl)hydrazones, J. Med. Chem. 51 (2008) 4874–4880.
[49] L. Santana, E. Uriarte, H. González-Díaz, G. Zagotto, R. Soto-Otero, E. Mendez-Alvarez, A QSAR model for in silico screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins, J. Med. Chem. 49 (2006) 1149–1156.
[50] L. Santana, H. Gonzalez-Diaz, E. Quezada, E. Uriarte, M. Yanez, D. Vina, F. Orallo, Quantitative structure-activity relationship and complex network approach to monoamine oxidase a and B inhibitors, J. Med. Chem. 51 (2008) 6740–6751.
[51] N.T. Nguyen, D.M. Cook, L.A. Bero, The decision-making process of US Food and Drug Administration advisory committees on switches from prescription to over-the-counter status: a comparative case study, Clin. Ther. 28 (2006) 1231–1243.
[52] A. Duardo-Sanchez, G. Patlewicz, A. Lopez-Diaz, Current topics on software use in medicinal chemistry: intellectual property, taxes, and regulatory issues, Curr. Top. Med. Chem. 8 (2008) 1666–1675.
[53] H. González-Díaz, F. Prado-Prado, L.G. Pérez-Montoto, A. Duardo-Sánchez, A. López-Díaz, QSAR models for proteins of Parasitic Organisms, Plants and

human Guests: theory, applications, legal Protection, taxes, and regulatory issues, Curr. Proteomics 6 (2009) 214–227.
[54] H. González-Díaz, A. Duardo-Sanchez, F.M. Ubeira, F. Prado-Prado, L.G. Pérez-Montoto, R. Concu, G. Podda, B. Shen, Review of MARCH-INSIDE & complex networks prediction of drugs: ADMET, anti-parasite activity, Metabolizing Enzym. Cardiotoxicity Proteome Biomarkers Curr. Drug Metab. 11 (2010) 379–406.
[55] H. Gonzalez-Diaz, F. Romaris, A. Duardo-Sanchez, L.G. Perez-Montoto, F. Prado-Prado, G. Patlewicz, F.M. Ubeira, Predicting drugs and proteins in parasite infections with topological indices of complex networks: theoretical backgrounds, applications, and legal issues, Curr. Pharm. Des 16 (2010) 2737–2764.
[56] C. Yang, L.G. Valerio Jr., K.B. Arvidson, Computational toxicology approaches at the US Food and drug Administration, Altern. Lab. Anim. 37 (2009) 523–531.

3D MI-DRAGON: new model for reconstruction of US FDA drug-target network and theoretic-experimental studies of rasagiline derivatives inhibitors for AChE.

Francisco Prado-Prado [1*], Xerardo García-Mera [1], Manuel Escobar[1], Nerea Alonso[1], Olga Caamaño[1],Matilde Yañez [2], and Humberto González-Díaz [3]

[1] *Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela (USC), 15782, Spain.*
[2] *Department of Pharmacology, Faculty of Pharmacy, USC, 15782, Spain*
[3] *Department of Microbiology and Parasitology, Faculty of Pharmacy, USC, 15782, Spain*

Abstract. The Neurodegenerative diseases have been increasing in the last years. Many of the drugs candidates to be used in the treatment of neurodegenerative disease present specific 3D structural features. One important protein in this sense is the acetylcholinesterase (AChE); which is the target of many Alzheimer's dementia drugs. Consequently, the prediction of Drug-Proteins Interactions (DPIs/nDPIs) between new drugs candidates with specific 3D structure and targets it is of the major importance. For it, we can use Quantitative Structure-Activity Relationships (QSAR) models to carry out rational DPIs prediction. Unfortunately, many previous QSAR models developed to predict DPIs take into consideration only 2D structural information and codify the activity against only one target. To solve this problem we can develop one 3D multi-target QSAR (3D mt-QSAR) models. In this communication, we introduce the technique 3D MI-DRAGON a new predictor for DPIs based two different well-known software. We use the software MARCH-INSIDE (MI) and DRAGON to calculate 3D structural parameters for drugs and targets respectively. Both classes of 3D parameters were used as input to train Artificial Neuronal Network (ANN) algorithms using as benchmark dataset the complex network (CN) formed by all DPIs between US FDA approved drugs and their targets. The entire dataset was downloaded from Drug Bank. The best 3D mt-QSAR predictor found is one ANN of type Multi-Layer Perceptron (MLP) with profile MLP 37:37-24-1:1. This MLP classifies correctly 274 out of 321 DPIs (Sensitivity = 85.35%) and 1041 out of 1190 nDPIs (Specificity = 87.48%), corresponding to training Accuracy = 87.03%. We validated the model with external predicting series with Sensitivity = 84.16% (542/644 DPIs; Specificity = 87.51% (2039/2330 nDPIs) and Accuracy = 86.78%. The new CNs of DPIs reconstructed from US FDA can be used to explore large DPIs databases in order to discover both new drugs and/or targets. We carried out theoretic-experimental studies to illustrate the practical use of 3D MI-DRAGON. First, we reported the prediction and pharmacological assay of 22 different rasagiline derivatives with possible AChE inhibitory activity.

Keywords: Drug-Protein interaction complex networks; Protein Structure Networks; multi-target QSAR; Markov Model; AChE inhibitors

Corresponding authors: PRADO PRADO, F. (francisco.prado@usc.es), Faculty of Pharmacy, University of Santiago de Compostela 15782, Spain., Fax: +34-981 594912.

1. Introduction

Yildirim, *et al.* [1] have built a complex network (CN) of Drug-Protein Pairs (DPIs) with the form of a bipartite graph composed of all DPIs for all US Food and Drug Administration (US FDA) approved drugs and proteins linked by drug-target binary associations. The resulting CN connects most drugs into a highly interlinked giant component, with strong local clustering of drugs of similar types according to Anatomical Therapeutic Chemical classification. It was motivated due to the strong incentive to develop new methods able of predicting potential drug-target interactions complex networks (CNs) formed by DPIs [2]. For it, we can use Quantitative Structure-Activity Relationships (QSAR) models [5] to carry DPIs prediction. To solve this problem we can develop a 3D multi-target QSAR (3D mt-QSAR) models to predict DPIs [6]. One way to develop this class mt-QSAR is incorporating into the QSAR equation parameters of the structure of the target (protein, DNA, RNA, etc.) in addition to the structural parameters of the drug present in classic QSAR. Some of the more known software we can use to reach this goal are: DRAGON, CODESSA[7], MODES-LAB[8], TOMO-COMD[9], and MARCH-INSIDE (MI)[10]. The software DRAGON is one of the more complete calculating more than 1600 descriptors for drug structure including as zero- (0D) one- (1D), two- (2D), three-dimensional (3D) parameters.

Unfortunately several QSAR models are able to predict the activity of drugs against only one target and/or are unable to codify important 3D structural features. Speck-Planche, *et al.*[3, 4] have developed mt-QSAR for the design of multi-target inhibitors against chemokine receptors. This approach was focused on the construction of a mt-QSAR model for the classification and prediction of inhibitors chemokine receptors. For instance, very recently we have developed in a previous work a QSAR model base on the MARCH-INSIDE method to predict a large network of DTPs [11].This model was based on 2D structural parameters for drugs and 1D structural parameters for protein. After that we developed MIND-BEST [12] and NL MIND-BEST [13]. Both predictors are based on 3D structural parameters of proteins calculated with software MI but they used only 2D structural parameters of drugs (calculated also with MI). The accuracy of the MIND-BEST model found was 86.32% and NL MIND-BEST was Accuracy = 90.41%. However both models only use 2D parameters using MI software. After that, to improve and obtain better results we use the software MARCH-INSIDE (MI) to calculate 3D structural parameters for targets and the software DRAGON was used to calculated 2D molecular descriptors all drugs[14]. We introduce the technique 2D MI-DRAGON a new predictor for DPIs based on two different well-known software.

As was mentioned in the previous paragraph we can seek a QSAR predictor for DPIs using molecular descriptors of both drug and target. In this work, we introduce for first time 3D MI-DRAGON a new predictor for DPIs based on two different well-known software. We use the software MARCH-INSIDE (MI) to calculate 3D structural parameters for targets and the software DRAGON for 3D parameters of all DPIs present in the Drug Bank (US FDA benchmark dataset) [15-18]. Both classes of parameters were used as input of different Artificial Neuronal Network (ANN) algorithms to seek an accurate non-linear mt-QSAR predictor. 3D MI-DRAGON offers a good opportunity for fast-track calculation of all possible DPIs of one drug enabling us to re-construct large drug-target or DPIs Complex Networks (CNs). In this study, we reported the prediction and pharmacological assay of 22 different rasagiline derivatives with AChE inhibitory activity. The present work reports the

attempts to calculate within unified DPIs. All this can help to design new inhibitors of AChE. A very good 3D MI-DRAGON QSAR model was obtained, and the subsequent combined QSAR & CN analysis may become of major importance for the prediction of the activity of new compounds against different targets or the discovery of new targets. In this sense we reported an illustrative study that combines both experiment and theory to show how to use this model in practical situations. We reported the prediction and pharmacological assay of rasagiline derivatives with AChE inhibitory activity. In **Figure 1** we depict a flowchart with the main steps given in this work to train and validate the ANN classifier.

Figure 1 comes about here

2. Materials and Methods

2.1. Computational methods

2.1.1 MOPAC AM1 Optimization geometry method using CS CHEM 3D.

Molecular structures of all FDA drugs were generated with CHEM 3D Ultra (version 2005). The energy of each intermediate was then minimized using the semi-empirical MOPAC method with a minimum RMS gradient of 0.100, which specifies the convergence criteria for the gradient of the potential energy surface. The geometry of the molecules was optimized and the values of the quantum chemical descriptors of each compound were calculated using AM1. AM1 theory was used with a closed shell function. The MOPAC AM1 method was selected because it was a semi-empirical quantum chemical method and the computational time was much shorter than that needed by ab initio method.

2.1.2. MI-DRAGON technique

3D Parameters for drugs. The DRAGON software 4.0 [19] was utilized here to calculate the 3D parameters of drugs. It depends on whether they are computed from the chemical formula, substructure list representation, molecular graph or geometrical representation of the molecule, respectively [20, 21]. In this work, we calculated only GETAWAY 3D descriptors. We use these descriptors after optimized for use with 3D descriptors.

3D Parameters of proteins. In previous works we have predicted protein function based on different protein structural parameters derived from a Markov matrix that account for electrostatic interactions between aminoacid pairs in the 3D structure of the protein. One of the classes of parameters used was called the Shannon Entropy ${}^{T}\theta_k(R)$ of the markov matrix. These values are used here as inputs to describe information about the structure of the drug target proteins (T) in order to construct the mt-QSAR models for DTPs. The detailed explanation has been published before [22-30] and reviewed in detail more recently [31]. At follows we give the formula for ${}^{T}\theta_k(R)$ values and some general explanations:

$$ {}^{T}\theta_k(R) = -\sum_{j \in R}{}^{k} p_j(R) \cdot \log\left[{}^{k} p_j(R)\right] \quad (1) $$

Where, ${}^{k}p_i(R)$ values are the absolute probabilities with which the effect of the electrostatic interaction propagates from the amino acid i^{th} to other amino acids j^{th} next to it and returns to i^{th} after k-steps. These probabilities refer to: aminoacids considered isolated in the space (k = 0), interaction between aminoacids in direct contact (k = 1) or spatial (k > 1) indirect interactions between amino acids placed at a distance equal to *k*-times the cut-off distance ($r_{ij} = k \cdot r_{cut\text{-}off}$) in the residue network. Euclidean 3D space $r_3 = (x, y, z)$ coordinates of the C_α atoms of amino acids listed in protein PDB files. For calculation, all water molecules and metal ions were removed [32]. All calculations were carried out with our in-house software MARCH-INSIDE 2.0 [32]. For the calculation, the MARCH-INSIDE

software always uses the full matrix, never a sub-matrix, but the last summation term may run either for all amino acids or only for some specific protein regions (R) denoted as: *c* for core, *i* for inner, *m* for middle, and *s* for surface regions, respectively). Consequently, we can calculate different ${}^{T}\theta_k(R)$ for the amino acids contained in the regions (*c*, *i*, *m*, *s*, or *t*) and placed at a topological distance *k* each other within this orbit (*k* is the order) [22, 23, 33-35]. In this work, we have calculated altogether 5(types of regions) x 6(orders considered) = 30 ${}^{T}\theta_k(R)$ indices for each protein.

2.1.3 Statistical analysis. Let be ${}^{D}\theta_k(G)$ entropy descriptors molecular that codify information about drug structure and ${}^{T}\theta_k(R)$ entropy descriptors that codify information about drug target proteins; we attempt to develop a simple mt-QSAR model in the form of a linear classifier with the general formula:

$$S(DTP)_{pred} = \sum_{k=0}^{5} a_{G,k} \cdot {}^{D}\theta_k(G) + \sum_{k=0}^{5} b_{R,k} \cdot {}^{T}\theta_k(R) + c_0 \quad (2)$$

We used Linear Discriminating Analysis (LDA) to fit this discriminant function. The model deals with the classification of a compound set with or without affinity on different receptors. A dummy variable Affinity Class (AC) was used as input to codify the affinity. This variable indicates either high (AC = 1) or low (AC = 0) affinity of the drug by the receptor. $S(DTP)_{pred}$ or DTP affinity predicted score is the output of the model and it is a continuous dimensionless score that sorts compounds from low to high affinity to the target coinciding DTPs with higher values of $S(DTP)_{pred}$ and nDTPs with lowest values. In equation (6), *b* represents the coefficients of the classification function, determined by the LDA module of the STATISTICA 6.0 software package [36]. We used Forward Stepwise algorithm for a variable selection. The statistical significance of the LDA model was determined calculating the p-level (p) of error with Chi-square test. We also inspected the Specificity, Sensitivity, and total Accuracy to determine the quality-of-fit to data in training. Cases for training set were selected at random out of the cases in full dataset. The remnant cases were used to validate the model. The validation of the model was corroborated with these external prediction series; these cases were never used to train the model. The ration between training/validation set was 2/1 approximately. This procedure to select training and validation sets is largely known and used to train QSAR models [37-43].

2.1.4 ANN analysis. The non-linear mt-QSAR model was constructed using ANN analysis. All models trained were carried out in STATISTICA 6.0 [36]. In so doing, we used a very simple type of ANN called Three Layers Perceptron (MLP-3) to fit this discriminant function. The model deals with the classification of a compound set with or without affinity on different receptors. A dummy variable Affinity Class (AC) was used as input to codify the affinity. This variable indicates either high (AC = 1) or low (AC = 0) affinity of the drug by the receptor. $S(DTP)_{pred}$ or DTP affinity predicted score is the output of the model and it is a continuous dimensionless score that sorts compounds from low to high affinity to the target coinciding DTPs with higher values of $S(DTP)_{pred}$ and nDTPs with lowest values. In equation (2), *b* represents the coefficients of the LNN classification function, determined by the ANN module of the STATISTICA 6.0 software package [36]. We used Forward Stepwise algorithm for a variable selection.

In addition, we can explore more complicated non-linear ANNs in order to improve the accuracy of the classifier. We processed our data with different ANNs looking for a better model. Four types of ANNs were used, namely, Probabilistic Neural Network (PNN), Radial Basic Function (RBF), Linear Neural Network (LNN), and Four Layer Perceptron

(MLP-4)[44, 45]. The quality of all the ANNs (linear or non-linear) was determined calculating values of Specificity, Sensitivity, and total Accuracy to determine the quality-of-fit to data in training. The validation of the model was corroborated with external prediction series. We also reported ROC-curve analysis (ROC curve can be used to select an optimum decision) for both training and validation series [44, 46].

2.1.5 Data set. The data set was formed by a set of marketed DPIs with known affinity of drugs by targets. This dataset is the same benchmark data used in previous works [1, 5, 12, 13] in this area and contains all drugs approved by the US FDA. We download this dataset from the public resource called Drug Bank [12, 13, 16-18]. The data set was formed for more than 519 drugs with their respectively 336 targets. Subsequently, we were able to collect above 4485 cases (drug-protein interactions) instead of 519 x 336 cases. In addition the data set was used to develop ANN models to performance the model. The names or codes for all compounds are depicted in **Table 1SM** of the supplementary material, due to space constraints, as well as the references consulted to compile the data in this table.

2.1.6 Complex network construction. We construct a DPIs network in order to achieve the drug and protein affinity with a network approach. Generally in this network, one node may represent a drug or a target. On the other hand, the edges represents the DPIs; express relationships between pairs of drugs with their targets [1]. Anyhow, the nodes representing targets may be of at least two types. In almost all cases reported up to date each target is represented only once in the network. In this class of "static" DPIs network the target is depicted by the node corresponding to the X-ray structure of itself. In this work, we build in total two complex networks. First, we constructed the DPIs networks for the observed data and second, DPIs network predicted by the model. The common steps to construct these networks are: First, using the Excel software in a column we introduce all the proteins, the drugs used quotation marks in our database. Then in another column lists all the cases. At the beginning of this column puts the total number of vertices, there are currently two columns of the name of drug and protein and their corresponding number of vertices. After, at the end of the columns are placed bows in the first column put the number of vertices for the drug and in another column corresponding to the protein. Then, the file was saved as a .txt format file. After we had renamed the .txt file as a .net file we read it with the CentiBin software [47, 48]. Finally, using CentiBin we can not only represent the network but also highlight all drugs and targets (nodes) connected by a specific edge or link (DPI). Using this software we can calculate vertex centralities to analyze the relationships between drug targets.

2.2. Illustrative experiments

2.2.1. Synthesis of Rasagiline derivatives.

Synthesis. Synthesis of compounds **1-22** has been previously reported by us [5, 12], see **Figure 2**.

Figure 2 comes about here

2.2.2 Determinations of cholinesterases activities

The cholinesterase assay method of Ellman was used to determine the in vitro cholinesterase activity [49]. The activity was measured by increase in absorbance at 412 nm due to the yellow color produced from the reaction of acetylthiocholine iodide with the dithiobisnitrobenzoate (DTNB) ion.

Acetylcholinesterase from human erythrocytes, acetylcholinesterase recombinant expressed in HEK 293 cells and butyrylcholinesterase from human serum was obtained from Sigma.

2.2.3 Experimental conditions and kinetics. Enzyme activity was measured using a FLUOstar Optima microplate reader. The assay medium contained phosphate buffer, pH 8.0, 20 mM DTNB, 0.01 U/ml of enzyme and 0.75 μM substrate (acetylthiocholine iodide or butyrylthiocholine iodide). The activity was determined by measuring the increase in absorbance at 412 nm at 1 min intervals for 10 min at 37 °C. In dose-dependent inhibition studies, the substrate was added to the assay medium containing enzyme, buffer, and DTNB with inhibitor after 10 min of incubation time. All experiments were carried out in duplicate and expressed as mean ± SEM. The relative activity is expressed as percentage ratio of enzyme activity in the absence of inhibitor, see **Table 1**.

Table 1 comes about here

3. Results

3.1. DPIs QSAR predictive models

3.1.1 LDA model. Common physicochemical properties like entropy have been demonstrated to be useful on protein QSAR [50, 51]. We used these properties as input of our model in addition to drug molecular descriptors. The present is the first mt-QSAR model combining DRAGON and MI to predict the probability with which occur DPIs between a drug and a protein. This type of models lie within the frontiers between classic QSAR for drugs and protein QSAR [33]. Some applications for the present model are the prediction of new drugs, new protein receptors or drug targets, and drug binding sites. Detailed information on the compounds, predicted classification, and probability of affinity on different receptors of the drugs used to seek the model appears in **Table 1SM** of the supplementary material. Based on the algorithms described in materials and methods the best linear model found was the following:

$$S(DTP)_{pred} = 11.01 \cdot d_1 - 36.37 \cdot d_2 - 12.77 \cdot d_3 - 9.34 \cdot d_4 - 52.17 \cdot d_5 + 1.62 \cdot d_6 - 1.65 \cdot d_7 - 0.11 \cdot d_8 - 0.10 \cdot d_9 + 0.25 d_{10} + 2.48 \quad (3)$$

$$N = 4485 \qquad \chi^2 = 919.2988 \qquad p - level < 0.001$$

Table 2 comes about here

The nomenclature used in the descriptors of the equation is found in **Table 2**. In this equation, N is the number of cases, χ^2 is the Chi-square and p is the level of error. This model, with 10 variables, classifies correctly 256 out of 321 DPIs (Sensitivity of 79.75%) and 1014 out of 1190 nDPIs (Specificity of 85.21%). Overall training Accuracy was 84.05%. The validation of the model was carried out by means of external predicting series. The model classifies correctly 498 out of 644 DPIs (77.33%) and 2000 out of 2330 nDPIs (85.84%) in validation series. Accuracy for validation series (predictability) was 83.99%. These results (**Table 3**) indicate that we developed an accurate model according to previous reports on the use of LDA in QSAR [52, 53].

Table 3 comes about here

3.3.2 3D MI-DRAGON ANN model. The previous model show good results with a relatively small number of parameters (10 parameters) and a linear equation. However, as result of the previous section we decided to carry out an ANN analysis to seek a better model using a non-linear method. Four types of ANNs were used, namely, Probabilistic Neural Network (PNN), Radial Basic Function (RBF), Three Layers Perceptron (MLP-3), and Four Layer Perceptron (MLP-4). See, previous works about the use of these ANNs in protein QSAR [5, 13]. The **Figure 3** depicts the networks topology for some of the ANN models tested. In general, at least one ANN of every type tested was statically significant.

However, one must note that the profiles of each network indicate that many of these are highly non-linear and complicated models.

Figure 3 comes about here

Models using ANN-QSAR has been demonstrated before; see, for instance, the works of Fernandez and Caballero [54, 55]. We compare different types of networks to obtain a better model. In **Table 3** we show the classification matrix of the different networks. The profiles of networks tested were RBF 1:1-1-1:1 with only one variable; LNN 227:227-1:1, which present many variables, and PNN 227:227-14797-2-2:1, which has a very high number of hidden neurons, see **Table 3.** After that, the simpler but more accurate ANN model found was an MLP (MLP 37:37-24-1:1) with training Accuracy = 87.03 %. This was selected as the best network found because it presents both high accuracy and an adequate number of variables accounting for features relevant for DPIs. This ANN presents 37 inputs variables (24 d_k + 13 Θ_m). This leads to 37 neurons in first or input layer (I), 24 neurons in the second layer or first hidden layer (H1) and only one neuron (DPI prediction) in the output layer (O). We depict the ROC-curve for MLP 37:37-24-1:1 to show how reliable was the network model developed, see **Figure 4**. Notably, the model presented had a ROC curve higher than 0.5. The model presented an area greater than 0.92. From now on we call the ANN MLP 37:37-24-1:1 as the 3D MI DRAGON predictor.

Figure 4 comes about here

3.3.2.1 3D MI-DRAGON assembly of CNs for DPIs., The construction of multi-protein CNs that incorporates protein affinity profile for drugs or the same CNs for DPIs is relevant to drug and target screening. And is one application for this model. In order to recall the capacity of 3D MI-DRAGON to predict new CNs of DPIs we selected the same benchmark database used in previous works [5, 13, 14]; which includes US FDA approved drugs with their targets. With these goals in mind, we constructed again and manually curated the above-mentioned CN obtaining a graph with 855 vertices or nodes (drugs and proteins) and m = 1016 DPIs (edges). This CN of DPIs have D = 6.7; average topological distances D_{ij} between all pairs of nodes. The same as before, we constructed a new CN of DPIs but connecting only pairs of nodes with DPIs predicted by 3D MI-DRAGON. In so doing, we obtained a value of D = 7.2 and m = 1256 DPIs. In **Figure 5** we illustrated visually both CNs (observed and predicted).

Figure 5 comes about here

In first instance, we compare this predicted network (3D MI-DRAGON) with 2D MI-DRAGON predicted network[14]. We compare to observe the similar or dissimilar topology (connectivity patterns structure) between them. Measuring in terms of TIs such as: number of nodes (n), number of edges (m), Wiener index (W), diameter (D), the Randic connectivity index (Xr), topological distance (Dist), network average values for radiality (R), node degree (δ), eccentricity (E). In **Table 4**, we observe all the TIs are similar excepting n, m and w. That means both CNs have a high similarity between them. These results are very interesting, because our 3D MI-DRAGON model present similar results to the 2D MI-DRAGON model, which results have been published successfully before.

Table 4 comes about here

To see how reliable and valid is our model. Not only compared to TIs to observe similarity between both predicted networks, but we study the centrality analysis of given networks too. This type of drug screening and drug target discovery is the calculation of those nodes (drugs or proteins) which are more relevant or important (central) in the graph. For it we can use numerical parameters that quantify the importance of a node in a graph which are

called node centralities C_t of type t [56]. These nodes identification using node centralities may help us to identify the more relevant drugs or proteins in analogy to similar procedures developed for PINs; networks of Protein-Protein Interactions (PPIs) [57]. In **Table 5** we show the predicted results of both node degree centrality (C_δ) and closeness centrality (C_{clo}) for proteins and drugs present in the database and compare with the predicted results of 2D-MI-DRAGON model. The parameter C_δ measures the local importance of a node by counting the number of nodes directly attached to him [57]. Conversely, C_{clo} measures the global importance of a node in a CN by taking in consideration the inverse of the sum of D_{ij} ($C_{clo} = 1/\Sigma D_{ij}$) [58]. Consequently, the higher C_δ the higher is the local importance of the node but the higher C_{clo} the lower is the global importance of the node. For instance, the protein 1HA2 is one important protein both locally and globally in this CNs with lower $C_{clo} > 4$ and a $C_\delta = 26$. It means that this protein is both locally and globally important because it is the target of many drugs (high C_δ). This result is similar to obtained by 2D MI-DRAGON model. Another interesting result was simvastatin. Simvastatin is a hypolipidemic drug used to control elevated cholesterol, or hypercholesterolemia. It is a member of the statin class of pharmaceuticals. The primary use of simvastatin is for the treatment of dyslipidemia and the prevention of cardiovascular disease [59, 60]. Depending on our aims the more important nodes in pharmacological terms not necessarily have to be the more central in the graph (those with higher C_δ and lower C_{clo}), see **Table 5.** We show in this example, our model predicts efficiently. We found that the 3D MI-DRAGON model shows very similar results to the previous model, which has been published with excellent results. In **Table 2SM** of supplementary material we show all node degrees and closeness results.

Table 5 comes about here

3.2. Theoretic-Experimental Study using 3D MI-DRAGON predictor

Finally, we illustrated in one theoretic-experimental study the practical use of 3D MI-DRAGON. We reported the prediction, synthesis, and pharmacological assay of 20 different rasagiline derivatives with AChE inhibitory activity.

3.2.1 3D MI-DRAGON prediction of rasagiline derivatives *vs.* AChE. In this *in silico* experiment we used 3D MI-DRAGON to predict the interaction of the rasagiline derivatives with respect to AChE. For it, we downloaded the 3D structure of AChE protein with PDB ID 1EEA and calculated their structural parameters with MI. We also generated the SMILE codes for these compounds and we use MOPAC AM1 Optimization geometry method for these compounds for calculated their 3D structural parameters with DRAGON. After that, we predicted their propensity to undergo DPIs with AChE using as inputs for the 3D MI-DRAGON predictor the structural parameters of both the drugs and the protein. In **Table 6** we confront the results obtained using this model and the outcomes of the pharmacological assay. No compounds are selective inhibitors of AChE, which is why we used as control galantamine for AChE was. We consider the observed class for active compounds OC = 1 if compound $IC_{50} < 10$ μM this cutoff is in the similar range than other used in previous works [61, 62]. As we can see in this table all the compounds rasagiline derivatives present some activity, But none of these compounds have inhibitory activity in the pharmacological assays. All of our compounds in the pharmacological assay (OC = 0) were inactive. 3D MI-DRAGON predicted as inactive all compounds, excepting 3. The model classified correctly 19 of 22 compounds tested (86.36%). In this test, our model was compared with pharmacological testing of 22 compounds synthesized by us. And we can observe the effectiveness of our model with experimental data. Also, we note that the

model predicts all compounds tested as inactive, this is important because the model allows to discriminate between active and inactive compounds. However, some compounds were not tested by pharmacological assay, that compounds were predicted as inactive using 3D MI-DRAGON model. We discarded pharmaceutical assays of these compounds; because we consider our model reliable. This kind of model can be used to saves efforts and money to perform the pharmacological tests. This is a good example of how reliable is the MI DRAGON 3D model.

Table 6 comes about here

3.2.2 3D MI-DRAGON complex network of rasagiline derivatives *vs.* US FDA proteins. An additional use of 3D MI-DRAGON was to carry out the "*in silico*" or virtual screening of the new compounds with respect to all other targets previously approved by US FDA [14, 63]. It may help to found new targets for these drugs or discard possible toxicological effects depending on the other targets predicted and/or discarded for these compounds. This type of experiment is of the major importance due to the cost in terms of animal sacrifice, time, materials and human resources of the experimental assay of all compounds against all these targets, see recent reviews by Duardo-Sanchez *et al.* [64-67]. In fact, over a decade, the US FDA has been engaged in the applied research, development, and evaluation of computational toxicology methods used to support the safety evaluation of a diverse set of regulated products. The basis for evaluating computational toxicology methods is multi-factorial, including the potential for increased efficiency, reduction in the numbers of animals used, lower costs, and the need to explore emerging technologies that support the goals of the US FDA's Critical Path Initiative (e.g. to make decision support information available early in the drug review process)[68].

In this experiment, we downloaded the 3D structure of all proteins that are targets of US FDA approved drugs. Next, we calculated the structural parameters of all these proteins with MI. We also generated the SMILE codes for these compounds and and we use MOPAC AM1 Optimization geometry method for these compounds for calculated their 3D structural parameters with DRAGON. After that, we predicted their propensity to undergo DPIs with all US FDA proteins using as inputs for the 3D MI-DRAGON predictor the structural parameters of both the drugs and proteins. We predicted all proteins in FDA dataset *vs.* the 22 rasagiline derivatives. We found that most of 22 derivatives were predicted as non-active (low DPIs scores) against most proteins in the FDA database. Consequently, 3D MI-DRAGON predicts a high selectivity of rasagiline derivatives as AChE inhibitors. We can reach this goal because the model predicts these compounds as non-active with respect to most proteins that are targets of FDA drugs.

Using these results, we constructed a DP-CN for rasagiline derivatives and the FDA dataset (see **Figure 6**). As a result we obtained a CN with 87 nodes (FDA drugs, proteins, or rasagiline derivatives) and 166 DP (edges, DTPs). As In this network we can see that protein 1EEA (AChE) is predicted to interacts with compound 3, this protein is an AChE target [69]. These results are good because they agree with the experimental results presented in this paper where the compound 3 show low AChE activity. The use of such complex networks can help us find and predict new drugs-protein interactions, and therefore find new drugs with improved biological activity and fewer side effects, especially in neural disease.

Figure 6 comes about here

4. Conclusions

The 3D MI-DRAGON predictor based on structural parameters of drugs calculated with DRAGON and parameters of proteins calculated with MI. It is possible to seek excellent predictors for DPIs using as input structural parameters of drugs and proteins calculated with different programs and combined with ANN models. Combining **MARCH-INSIDE** and DRAGON approach and ANN is possible to seek one mt-QSAR classifier to predict with Accuracy > 85% the probability of drugs to bind more than 500 different drug target proteins approved by FDA of USA. 3D MI-DRAGON predictor is also useful to assemble CNs of DPIs. These CNs computationally assemble offer an alternative to discover new drugs or targets, and explore the selectivity of drugs. In this work, we exemplified these conclusions through the experimental-theoretical study of the AChE activity of new rasagiline derivatives.

5. Acknowledgments

Prado-Prado F. thanks sponsorships for research position at the University of Santiago de Compostela from *Angeles Alvariño*, Xunta de Galicia.

References

[1] M.A. Yildirim, K.I. Goh, M.E. Cusick, A.L. Barabasi, M. Vidal, Drug-target network, Nature biotechnology, 25 (2007) 1119-1126.
[2] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, M. Kanehisa, Prediction of drug-target interaction networks from the integration of chemical and genomic spaces, Bioinformatics, 24 (2008) i232-240.
[3] A. Speck-Planche, V.V. Kleandrova, F. Luan, M.N. Cordeiro, Chemoinformatics in Multi-Target Drug Discovery for Anti-Cancer Therapy: In Silico Design Of Potent And Versatile Anti-Brain Tumor Agents, Anticancer Agents Med Chem, (2011).
[4] A. Speck-Planche, V.V. Kleandrova, In silico design of multi-target inhibitors for C-C chemokine receptors using substructural descriptors, Mol Divers, (2011).
[5] F. Prado-Prado, X. Garcia-Mera, P. Abeijon, N. Alonso, O. Caamano, M. Yanez, T. Garate, M. Mezo, M. Gonzalez-Warleta, L. Muino, F.M. Ubeira, H. Gonzalez-Diaz, Using entropy of drug and protein graphs to predict FDA drug-target network: theoretic-experimental study of MAO inhibitors and hemoglobin peptides from Fasciola hepatica, European journal of medicinal chemistry, 46 (2011) 1074-1094.
[6] F.J. Prado-Prado, F. Borges, L.G. Perez-Montoto, H. Gonzalez-Diaz, Multi-target spectral moment: QSAR for antifungal drugs vs. different fungi species, Eur J Med Chem, 44 (2009) 4051-4056.
[7] A.M. Helguera, R.D. Combes, M.P. Gonzalez, M.N. Cordeiro, Applications of 2D descriptors in drug design: a DRAGON tale, Curr Top Med Chem, 8 (2008) 1628-1655.
[8] E. Estrada, E. Molina, D. Nodarse, E. Uriarte, Structural contributions of substrates to their binding to P-Glycoprotein. A TOPS-MODE approach, Curr Pharm Des, 16 (2010) 2676-2709.
[9] Y. Marrero-Ponce, G.M. Casanola-Martin, M.T. Khan, F. Torrens, A. Rescigno, C. Abad, Ligand-based computer-aided discovery of tyrosinase inhibitors. Applications of the TOMOCOMD-CARDD method to the elucidation of new compounds, Curr Pharm Des, 16 (2010) 2601-2624.
[10] H. Gonzalez-Diaz, A. Duardo-Sanchez, F.M. Ubeira, F. Prado-Prado, L.G. Perez-Montoto, R. Concu, G. Podda, B. Shen, Review of MARCH-INSIDE & complex networks prediction of drugs: ADMET, anti-parasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers, Curr Drug Metab, 11 (2010) 379-406.
[11] D. Vina, E. Uriarte, F. Orallo, H. Gonzalez-Diaz, Alignment-free prediction of a drug-target complex network based on parameters of drug connectivity and protein sequence of receptors, Mol Pharm, 6 (2009) 825-835.
[12] H. Gonzalez-Diaz, F. Prado-Prado, X. Garcia-Mera, N. Alonso, P. Abeijon, O. Caamano, M. Yanez, C.R. Munteanu, A. Pazos, M.A. Dea-Ayuela, M.T. Gomez-Munoz, M.M. Garijo, J. Sansano, F.M. Ubeira, MIND-BEST: Web server for drugs and target discovery; design, synthesis, and assay of MAO-B inhibitors and theoretical-experimental study of G3PDH protein from Trichomonas gallinae, Journal of proteome research, 10 (2011) 1698-1718.
[13] H. Gonzalez-Diaz, F. Prado-Prado, E. Sobarzo-Sanchez, M. Haddad, S. Maurel Chevalley, A. Valentin, J. Quetin-Leclercq, M.A. Dea-Ayuela, M. Teresa Gomez-Munos, C.R. Munteanu, J. Jose Torres-Labandeira, X. Garcia-Mera, R.A. Tapia, F.M. Ubeira, NL MIND-BEST: a web server for ligands and proteins discovery--theoretic-experimental study of proteins of Giardia lamblia and new compounds active against Plasmodium falciparum, J Theor Biol, 276 (2011) 229-249.

[14] F. Prado-Prado, X. Garcia-Mera, M. Escobar, E. Sobarzo-Sanchez, M. Yanez, P. Riera-Fernandez, H. Gonzalez-Diaz, 2D MI-DRAGON: a new predictor for protein-ligands interactions and theoretic-experimental studies of US FDA drug-target network, oxoisoaporphine inhibitors for MAO-A and human parasite proteins, European journal of medicinal chemistry, 46 (2011) 5838-5851.
[15] J. Kirchmair, P. Markt, S. Distinto, D. Schuster, G.M. Spitzer, K.R. Liedl, T. Langer, G. Wolber, The Protein Data Bank (PDB), its related services and software tools as key components for in silico guided drug discovery, J Med Chem, 51 (2008) 7021-7040.
[16] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A.C. Guo, D.S. Wishart, DrugBank 3.0: a comprehensive resource for 'omics' research on drugs, Nucleic Acids Res, 39 (2011) D1035-1041.
[17] D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, DrugBank: a knowledgebase for drugs, drug actions and drug targets, Nucleic Acids Res, 36 (2008) D901-906.
[18] D.S. Wishart, C. Knox, A.C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, DrugBank: a comprehensive resource for in silico drug discovery and exploration, Nucleic Acids Res, 34 (2006) D668-672.
[19] Talete srl, DRAGON for Windows (Software for Molecular Descriptor Calculations), in, 2005.
[20] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors., 2000.
[21] E. Papa, F. Villa, P. Gramatica, Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in Pimephales promelas (fathead minnow), J Chem Inf Model, 45 (2005) 1256-1266.
[22] H. González-Díaz, Y. Pérez-Castillo, G. Podda, E. Uriarte, Computational Chemistry Comparison of Stable/Nonstable Protein Mutants Classification Models Based on 3D and Topological Indices, Journal of Computational Chemistry, 28 (2007) 1990-1995.
[23] H. Gonzalez-Diaz, L. Saiz-Urra, R. Molina, Y. Gonzalez-Diaz, A. Sanchez-Gonzalez, Computational chemistry approach to protein kinase recognition using 3D stochastic van der Waals spectral moments, J. Comput. Chem., 28 (2007) 1042-1048.
[24] G. Aguero-Chapin, A. Antunes, F.M. Ubeira, K.C. Chou, H. Gonzalez-Diaz, Comparative Study of Topological Indices of Macro/Supramolecular RNA Complex Networks, Journal of chemical information and modeling, 48 (2008) 2265–2277.
[25] M. Cruz-Monteagudo, C.R. Munteanu, F. Borges, M.N.D.S. Cordeiro, E. Uriarte, K.-C. Chou, H. González-Díaz, Stochastic molecular descriptors for polymers. 4. Study of complex mixtures with topological indices of mass spectra spiral and star networks: The blood proteome case, Polymer, 49 (2008) 5575-5587.
[26] M.A. Dea-Ayuela, Y. Perez-Castillo, A. Meneses-Marcel, F.M. Ubeira, F. Bolas-Fernandez, K.C. Chou, H. Gonzalez-Diaz, HP-Lattice QSAR for dynein proteins: experimental proteomics (2D-electrophoresis, mass spectrometry) and theoretic study of a Leishmania infantum sequence, Bioorg. Med. Chem., 16 (2008) 7770-7776.
[27] G. Aguero-Chapin, H. Gonzalez-Diaz, G. de la Riva, E. Rodriguez, A. Sanchez-Rodriguez, G. Podda, R.I. Vazquez-Padron, MMM-QSAR recognition of ribonucleases without alignment: comparison with an HMM model and isolation from Schizosaccharomyces pombe, prediction, and experimental assay of a new sequence, Journal of chemical information and modeling, 48 (2008) 434-448.
[28] G. Ferino, H. Gonzalez-Diaz, G. Delogu, G. Podda, E. Uriarte, Using spectral moments of spiral networks based on PSA/mass spectra outcomes to derive quantitative

proteome-disease relationships (QPDRs) and predicting prostate cancer, Biochem. Biophys. Res. Commun., 372 (2008) 320-325.
[29] H. Gonzalez-Diaz, M.A. Dea-Ayuela, L.G. Perez-Montoto, F.J. Prado-Prado, G. Aguero-Chapin, F. Bolas-Fernandez, R.I. Vazquez-Padron, F.M. Ubeira, QSAR for RNases and theoretic-experimental study of molecular diversity on peptide mass fingerprints of a new Leishmania infantum protein, Mol. Divers., (2009).
[30] G. Aguero-Chapin, J. Varona-Santos, G.A. de la Riva, A. Antunes, T. Gonzalez-Villa, E. Uriarte, H. Gonzalez-Diaz, Alignment-Free Prediction of Polygalacturonases with Pseudofolding Topological Indices: Experimental Isolation from Coffea arabica and Prediction of a New Sequence, Journal of proteome research, 8 (2009) 2122-2128.
[31] H. Gonzalez-Diaz, F. Prado-Prado, F.M. Ubeira, Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach, Curr Top Med Chem, 8 (2008) 1676-1690.
[32] H. González-Díaz, Y. González-Díaz, L. Santana, F.M. Ubeira, E. Uriarte, Proteomics, networks and connectivity indices, Proteomics, 8 (2008) 750-778.
[33] H. Gonzalez-Diaz, L. Saiz-Urra, R. Molina, L. Santana, E. Uriarte, A model for the recognition of protein kinases based on the entropy of 3D van der Waals interactions, Journal of proteome research, 6 (2007) 904-908.
[34] H. Gonzalez-Diaz, R. Molina, E. Uriarte, Recognition of stable protein mutants with 3D stochastic average electrostatic potentials, FEBS Lett., 579 (2005) 4297-4301.
[35] R. Concu, G. Podda, E. Uriarte, H. Gonzalez-Diaz, Computational chemistry study of 3D-structure-function relationships for enzymes based on Markov models for protein electrostatic, HINT, and van der Waals potentials, J. Comput. Chem., 30 (2009) 1510-1520.
[36] StatSoft.Inc., STATISTICA (data analysis software system), version 6.0, www.statsoft.com.Statsoft, Inc., in, 2002.
[37] G.M. Casanola-Martin, Y. Marrero-Ponce, M.T. Khan, S.B. Khan, F. Torrens, F. Perez-Jimenez, A. Rescigno, C. Abad, Bond-based 2D quadratic fingerprints in QSAR studies: virtual and in vitro tyrosinase inhibitory activity elucidation, Chem Biol Drug Des, 76 (2010) 538-545.
[38] J.A. Castillo-Garit, M.C. Vega, M. Rolon, Y. Marrero-Ponce, V.V. Kouznetsov, D.F. Torres, A. Gomez-Barrio, A.A. Bello, A. Montero, F. Torrens, F. Perez-Gimenez, Computational discovery of novel trypanosomicidal drug-like chemicals by using bond-based non-stochastic and stochastic quadratic maps and linear discriminant analysis, Eur J Pharm Sci, 39 (2010) 30-36.
[39] R. Gozalbes, F. Barbosa, E. Nicolai, D. Horvath, N. Froloff, Development and validation of a pharmacophore-based QSAR model for the prediction of CNS activity, ChemMedChem, 4 (2009) 204-209.
[40] Y. Marrero-Ponce, A. Meneses-Marcel, O.M. Rivera-Borroto, R. Garcia-Domenech, J.V. De Julian-Ortiz, A. Montero, J.A. Escario, A.G. Barrio, D.M. Pereira, J.J. Nogal, R. Grau, F. Torrens, C. Vogel, V.J. Aran, Bond-based linear indices in QSAR: computational discovery of novel anti-trichomonal compounds, J Comput Aided Mol Des, 22 (2008) 523-540.
[41] S.J. Patankar, P.C. Jurs, Classification of inhibitors of protein tyrosine phosphatase 1B using molecular structure based descriptors, J Chem Inf Comput Sci, 43 (2003) 885-899.
[42] M. Murcia-Soler, F. Perez-Gimenez, F.J. Garcia-March, M.T. Salabert-Salvador, W. Diaz-Villanueva, P. Medina-Casamayor, Discrimination and selection of new potential antibacterial compounds using simple topological descriptors, J Mol Graph Model, 21 (2003) 375-390.

[43] R.A. Cercos-del-Pozo, F. Perez-Gimenez, M.T. Salabert-Salvador, F.J. Garcia-March, Discrimination and molecular design of new theoretical hypolipaemic agents using the molecular connectivity functions, J Chem Inf Comput Sci, 40 (2000) 178-184.
[44] F.J. Prado-Prado, X. Garcia-Mera, H. Gonzalez-Diaz, Multi-target spectral moment QSAR versus ANN for antiparasitic drugs against different parasite species, Bioorg. Med. Chem., 18 (2010) 2225-2231.
[45] Y. Rodriguez-Soca, C.R. Munteanu, J. Dorado, A. Pazos, F.J. Prado-Prado, H. Gonzalez-Diaz, Trypano-PPI: a web server for prediction of unique targets in trypanosome proteome by using electrostatic parameters of protein-protein interactions, Journal of proteome research, 9 (2010) 1182-1190.
[46] F.J. Prado-Prado, H. Gonzalez-Diaz, L. Santana, E. Uriarte, Unified QSAR approach to antimicrobials. Part 2: predicting activity against more than 90 different species in order to halt antibacterial resistance, Bioorg. Med. Chem., 15 (2007) 897-902.
[47] B.H. Junker, D. Koschutzki, F. Schreiber, Exploration of biological network centralities with CentiBiN, BMC Bioinformatics, 7 (2006) 219.
[48] D. Koschützki, CentiBiN Version 1.4.2, in, 2006, pp. CentiBiN Version 1.4.2, Centralities in Biological Networks © 2004-2006 Dirk Koschützki Research Group Network Analysis, IPK Gatersleben, Germany.
[49] E. GL, C. KD, A.V. Jr, F.-S. RM, A new and rapid colorimetric determination of acetylcholinesterase activity, Biochemistry Pharmacology, 7 (1961) 88-95.
[50] O. Ivanciuc, N. Oezguen, V.S. Mathura, C.H. Schein, Y. Xu, W. Braun, Using property based sequence motifs and 3D modeling to determine structure and functional regions of proteins, Current medicinal chemistry, 11 (2004) 583-593.
[51] C.H. Schein, O. Ivanciuc, W. Braun, Common physical-chemical properties correlate with similar structure of the IgE epitopes of peanut allergens, Journal of agricultural and food chemistry, 53 (2005) 8752-8759.
[52] Y.M. Alvarez-Ginarte, Y. Marrero-Ponce, J.A. Ruiz-Garcia, L.A. Montero-Cabrera, J.M. Vega, P. Noheda Marin, R. Crespo-Otero, F.T. Zaragoza, R. Garcia-Domenech, Applying pattern recognition methods plus quantum and physico-chemical molecular descriptors to analyze the anabolic activity of structurally diverse steroids, Journal of Computational Chemistry, (2007).
[53] A.H. Morales, J.E. Rodríguez-Borges, X. García-Mera, F. Fernández, M.N. Dias-Sueiro-Cordeiro, Probing the Anticancer Activity of Nucleoside Analogues: A QSAR Model Approach Using an Internally Consistent Training Set, Journal of Medicinal Chemistry, 50 (2007) 1537-1545.
[54] M. Fernandez, J. Caballero, A. Tundidor-Camba, Linear and nonlinear QSAR study of N-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivatives as matrix metalloproteinase inhibitors, Bioorg Med Chem, 14 (2006) 4137-4150.
[55] J. Caballero, M. Fernandez, Linear and nonlinear modeling of antifungal activity of some heterocyclic ring derivatives using multiple linear regression and Bayesian-regularized neural networks, J Mol Model, 12 (2006) 168-181.
[56] H. Gonzalez-Diaz, Y. Gonzalez-Diaz, L. Santana, F.M. Ubeira, E. Uriarte, Proteomics, networks and connectivity indices, Proteomics, 8 (2008) 750-778.
[57] H. Jeong, S.P. Mason, A.L. Barabasi, Z.N. Oltvai, Lethality and centrality in protein networks, Nature, 411 (2001) 41-42.
[58] E. Estrada, Virtual identification of essential proteins within the protein interaction network of yeast, Proteomics, 6 (2006) 35-40.

[59] P.A. Todd, K.L. Goa, Simvastatin. A review of its pharmacological properties and therapeutic potential in hypercholesterolaemia, Drugs, 40 (1990) 583-607.
[60] O. Hernandez-Perera, D. Perez-Sala, J. Navarro-Antolin, R. Sanchez-Pascuala, G. Hernandez, C. Diaz, S. Lamas, Effects of the 3-hydroxy-3-methylglutaryl-CoA reductase inhibitors, atorvastatin and simvastatin, on the expression of endothelin-1 and endothelial nitric oxide synthase in vascular endothelial cells, J Clin Invest, 101 (1998) 2711-2719.
[61] L. Santana, E. Uriarte, H. González-Díaz, G. Zagotto, R. Soto-Otero, E. Mendez-Alvarez, A QSAR model for in silico screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins, Journal of Medicinal Chemistry, 49 (2006) 1149-1156.
[62] L. Santana, H. Gonzalez-Diaz, E. Quezada, E. Uriarte, M. Yanez, D. Vina, F. Orallo, Quantitative structure-activity relationship and complex network approach to monoamine oxidase a and B inhibitors, J. Med. Chem., 51 (2008) 6740-6751.
[63] N.T. Nguyen, D.M. Cook, L.A. Bero, The decision-making process of US Food and Drug Administration advisory committees on switches from prescription to over-the-counter status: a comparative case study, Clin Ther, 28 (2006) 1231-1243.
[64] A. Duardo-Sanchez, G. Patlewicz, A. Lopez-Diaz, Current topics on software use in medicinal chemistry: intellectual property, taxes, and regulatory issues, Curr Top Med Chem, 8 (2008) 1666-1675.
[65] H. González-Díaz, F. Prado-Prado, L.G. Pérez-Montoto, A. Duardo-Sánchez, A. López-Díaz, QSAR Models for Proteins of Parasitic Organisms, Plants and Human Guests: Theory, Applications, Legal Protection, Taxes, and Regulatory Issues, Curr Proteomics, 6 (2009) 214-227.
[66] H. González-Díaz, A. Duardo-Sanchez, F.M. Ubeira, F. Prado-Prado, L.G. Pérez-Montoto, R. Concu, G. Podda, B. Shen, Review of MARCH-INSIDE & Complex Networks prediction of Drugs: ADMET, Anti-parasite Activity, Metabolizing Enzymes and Cardiotoxicity Proteome Biomarkers Current Drug Metabolism, 11 (2010) 379-406.
[67] H. Gonzalez-Diaz, F. Romaris, A. Duardo-Sanchez, L.G. Perez-Montoto, F. Prado-Prado, G. Patlewicz, F.M. Ubeira, Predicting drugs and proteins in parasite infections with topological indices of complex networks: theoretical backgrounds, applications, and legal issues, Curr Pharm Des, 16 (2010) 2737-2764.
[68] C. Yang, L.G. Valerio, Jr., K.B. Arvidson, Computational toxicology approaches at the US Food and Drug Administration, Altern Lab Anim, 37 (2009) 523-531.
[69] M.L. Raves, M. Harel, Y.P. Pang, I. Silman, A.P. Kozikowski, J.L. Sussman, Structure of acetylcholinesterase complexed with the nootropic alkaloid, (-)-huperzine A, Nat Struct Biol, 4 (1997) 57-63.

Figure Legends:

Figure 1. Flowchart of all steps given in this work to develop the new model

Figure 2. Rasagiline derivatives used in this work

Figure 3. Generic Topology of ANN models trained in this work

Figure 4. ROC Curve for 3D MI-DRAGONGON predictor (red = train series, blue = validation series

Figure 5. Observed vs. Predicted drug-target complex networks

Figure 6. Complex network of rasagiline derivatives vs. US FDA proteins

Tables:

Table 1. Inhibitory activity of different rasagiline derivatives .

Compounds	hAChE (IC_{50} μM)		hAChE (IC_{50} μM)
1	>100 μM	12	>100 μM
2	>100 μM	13	No tested
3	>100 μM	14	No tested
4	No tested	15	No tested
5	**	16	No tested
6	No tested	17	**
7	**	18	**
8	No tested	19	**
9	**	20	**
10	>100 μM	21	**
11	>100 μM	22	**
Galantamine	1.43 ± 0.03[a]		
Eserine	151.40 ± 5.63 nM		
Tacrine	130,90 ± 6,83 nM		

Each IC_{50} value is the mean ± S.E.M. from five experiments.

Table 2. Detailed list of the symbols and description for all parameters present in the model.

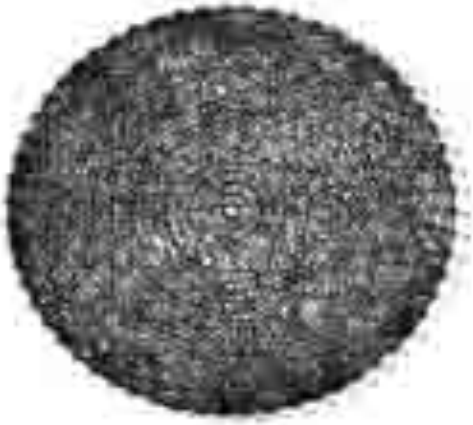
Original Descriptor	Descriptor name	Code ID
H7v	H autocorrelation of lag 7 / weighted by atomic van der Waals volumes	d_1
HATS5v	leverage-weighted autocorrelation of lag 5 / weighted by atomic van der Waals volumes	d2
HATS4e	leverage-weighted autocorrelation of lag 4 / weighted by atomic Sanderson electronegativities	d3
HATS6e	leverage-weighted autocorrelation of lag 6 / weighted by atomic Sanderson electronegativities	d4
R5e+	R maximal autocorrelation of lag 5 / weighted by atomic Sanderson electronegativities	d5
${}^{T}\theta_4(core)$	Entropy of all aminoacids placed in the core region and all the neighbors at distance k ≤ 4	d6
${}^{T}\theta_5(core)$	Entropy of all aminoacids placed in the core region and all the neighbors at distance k ≤ 5	d7
${}^{T}\theta_5(inner)$	Entropy of all aminoacids placed in the inner region and all the neighbors at distance k ≤ 5	d8
${}^{T}\theta_2(middle)$	Entropy of all aminoacids placed in the middle region and all the neighbors at distance k ≤ 2	p1
${}^{T}\theta_0(surface)$	Entropy of all aminoacids placed in the surface region and all the neighbors at distance k ≤ 0	d9

Table 3. Comparison of LDA and different ANNs classification models.

Model profile	Class	Train			Stat. Par.	Validation		
		%	DPIs	nDPIs		%	DPIs	nDPIs
MI DRAGON 3D	DPIs	85.36	274	47	Sn	84.16	542	102
MLP	nDPIs	87.48	149	1041	Sp	87.51	291	2039
37:37-24-1:1	Total	87.03			Ac	86.79		
LDA[a]	DPIs	79.75	256	65	Sn	77.33	498	146
10:10-1:1	nDPIs	85.21	176	1014	Sp	85.84	330	2000
	Total	84.05			Ac	83.99		
PNN	DPIs	0	0	644	Sn	0	0	321
227:227-14797-2-2:1	nDPIs	100	0	2346	Sp	100	0	1174
	Total	78.46			Ac	78.53		
RBF	DPIs	47.05	303	341	Sn	52.65	169	152
1:1-1-1:1	nDPIs	56.01	1032	1314	Sp	54.86	530	644
	Total	54.08			Ac	54.38		
LNN	DPIs	53.73	346	298	Sn	45.79	147	174
227:227-1:1	nDPIs	32.05	1594	752	Sp	31.52	804	370
	Total	36.72			Ac	34.58		

DPIs: Drug-Target Pairs for compounds with high affinity; nDPIs: Drug-Target Pair for compounds with non-affinity; Stat. is statistics, Par. is parameter

Table 4 Comparison 3D MI-DRAGON versus 2D MI-DRAGON.

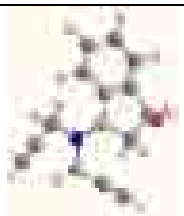
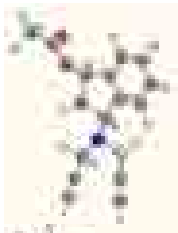
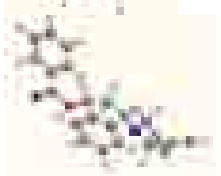
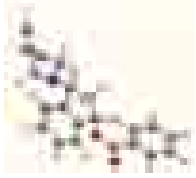
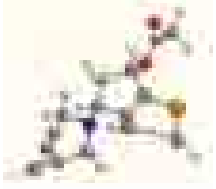
2D MI-DRAGON	Value	TIs	Value	3D MI-DRAGON
	706	n	59	
	907	m	631	
	1826812	W	2057954	
	18	D	19	
	255.09	Xr	266.39	
	2.49	δ	2.44	
	6.7	Dist	7.2	
	0.078	E	0.083	
	12.43	R	11.39	

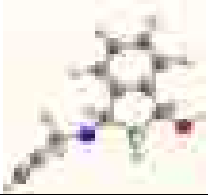
[a]The TIs used are: number of nodes (n), number of edges (m), Wiener index (W), diameter (D), the Randic connectivity index (Xr), topological distance (Dist), network average values for radiality (R), node degree (δ), eccentricity (E).

Table 5. Results of node degree (C_δ) and closeness centrality (C_{clo}) for 20 proteins and drugs.

Drug/PDB	C_δ 2D-MI-DRAGON	C_δ 3D MI-DRAGON	Drug/PDB	C_{clo} 2D-MI-DRAGON	C_{clo} 3D MI-DRAGON
1HA2	44	26	1HA2	4.80	3.54
1BNA	36	40	Simvastatin	4.22	4.11
NADH	35	33	Gliclazide	4.17	3.48
1R5K	27	29	Saquinavir	4.16	3.46
Simvastatin	18	17	1BNA	4.15	3.46
1EMI	16	21	Cefalotin	4.13	2.98
1CZM	14	17	Atorvastatin	4.09	3.44
1NHZ	14	14	1A8M	4.09	3.75
1MO8	14	14	1XF0	4.07	3.09
1UZF	13	13	Estrone	4.06	2.61
1SQN	13	13	Ketoprofen	4.02	2.86
1T9N	13	12	Testosterone	4.02	2.42
1BYW	11	15	1TZI	4.02	3.77
1VRU	11	10	1KED	4.00	3.16
1E3G	11	10	Captopril	3.99	3.49
Atorvastatin	11	10	Liothyronine	3.97	3.21
1ZNC	10	11	Diflunisal	3.96	3.20
1ODW	10	10	Halothane	3.95	3.11
Pyridoxal Phosphate	9	9	Digitoxin	3.94	3.45
1HWL	9	9	Pyridoxine	3.94	3.08

Table 6. Prediction of rasagiline derivatives with 3D MI-DRAGON predictor

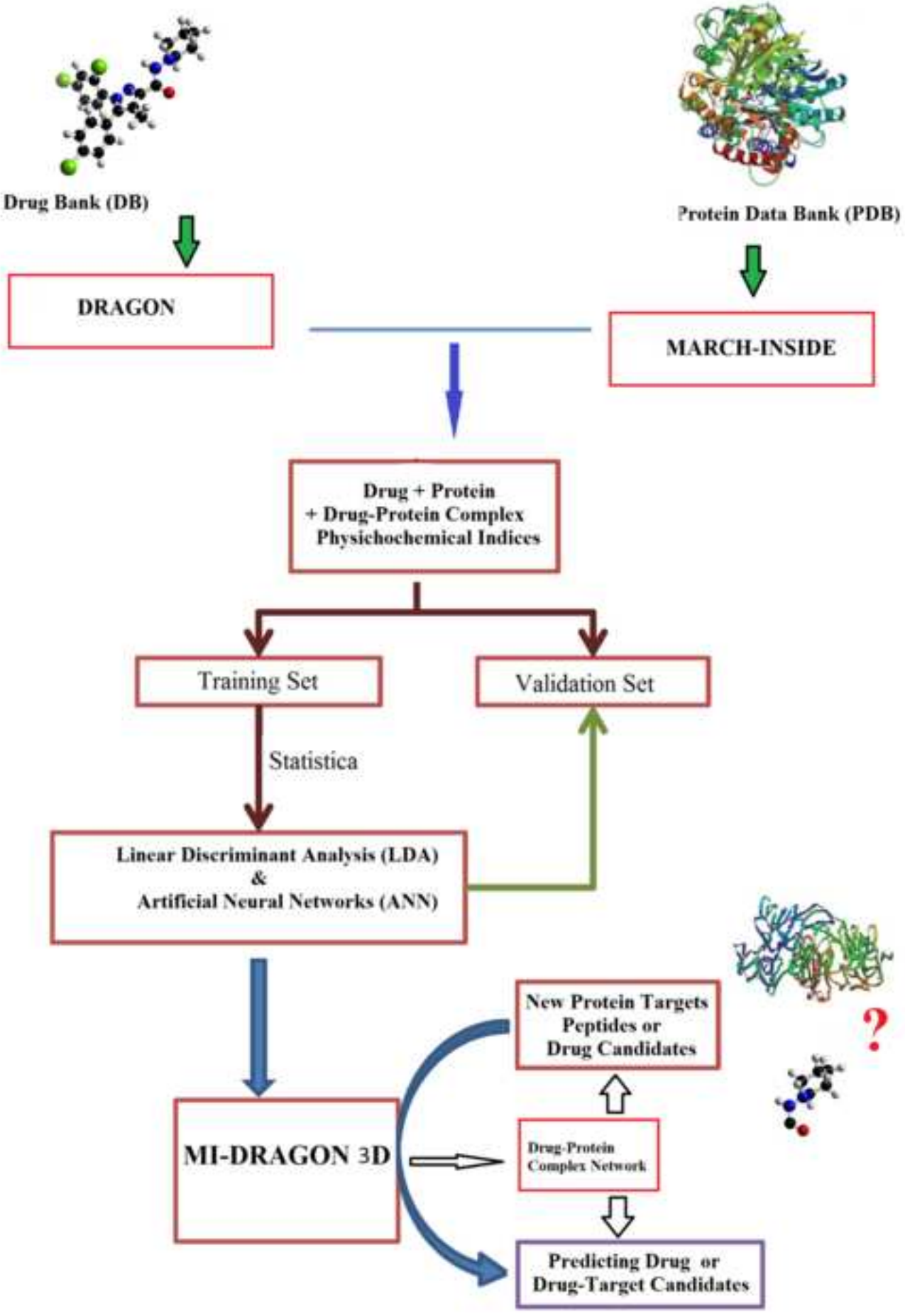
DRUG	OC	PC	Score	Structure	DRUG	OC	PC	Score	Structure
1	0	0	0.95		12	0	0	0.63	
2	0	0	0.95		13	0	0	1.00	
3	0	0	0.86		14	0	0	0.87	
4	0	0	0.95		15	0	0	1.00	
5	0	1	0.52		16	0	0	0.87	
6	0	0	0.95		17	0	0	1.00	
7	0	1	0.52		18	0	0	1.00	

8	0	0	0.88		19	0	0	1.00	
9	0	0	0.89		20	0	0	1.00	
10	0	0	0.75		21	0	0	0.97	
11	0	1	0.27		22	0	0	0.97	

OC = Observed class; PC = Predicted class

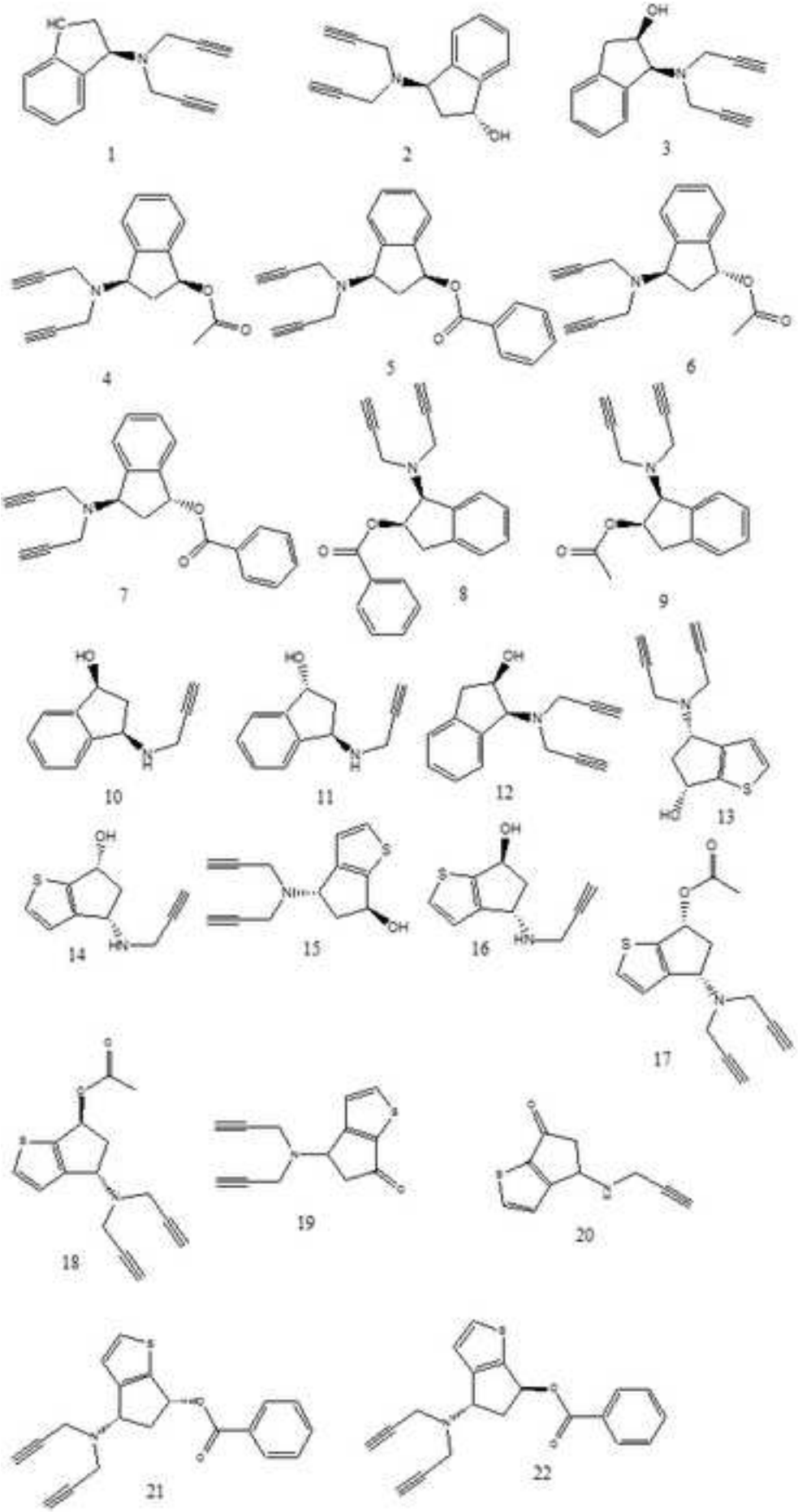
Figure(s)

Click here to download high resolution image



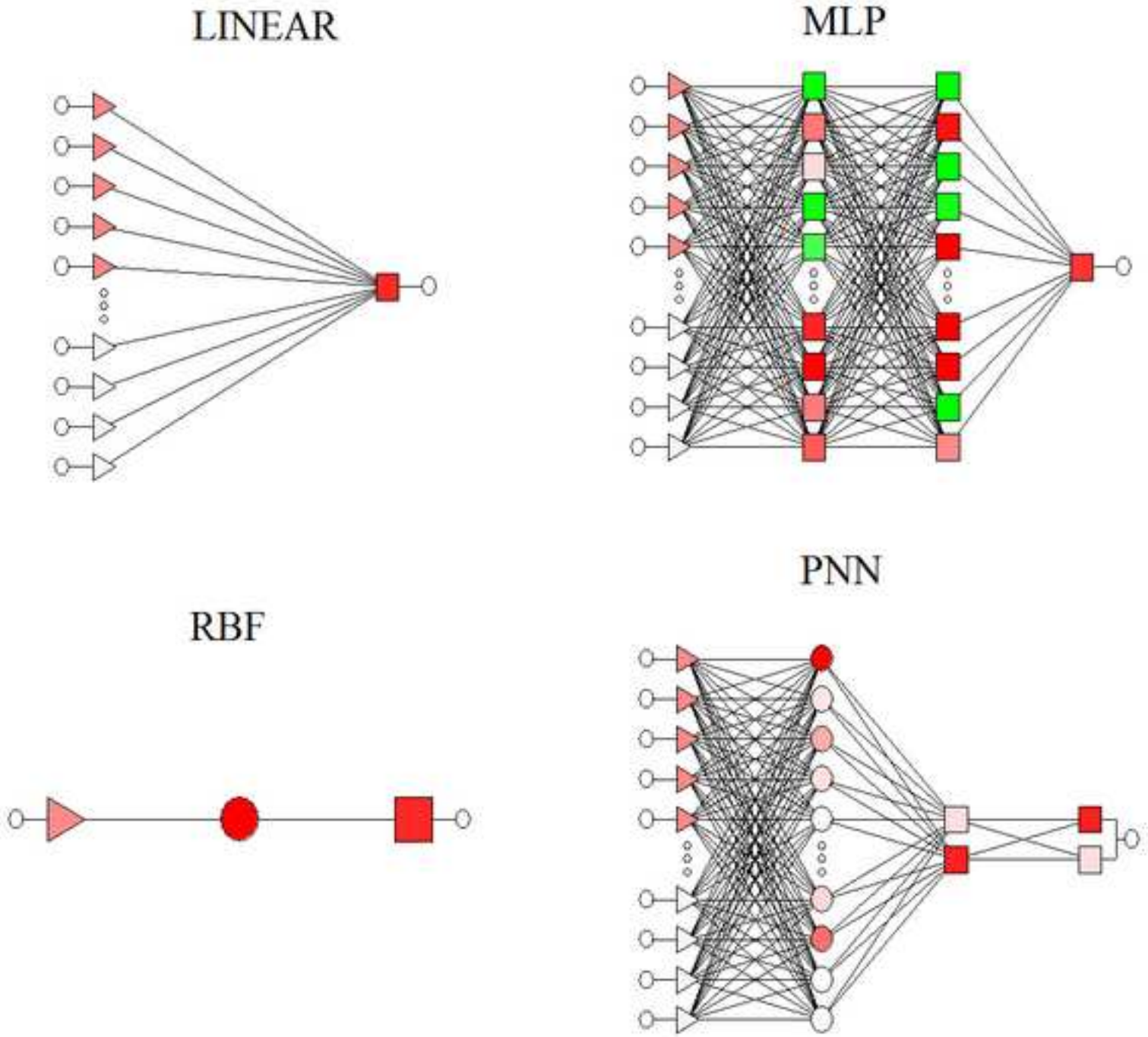
Figure(s)

Click here to download high resolution image



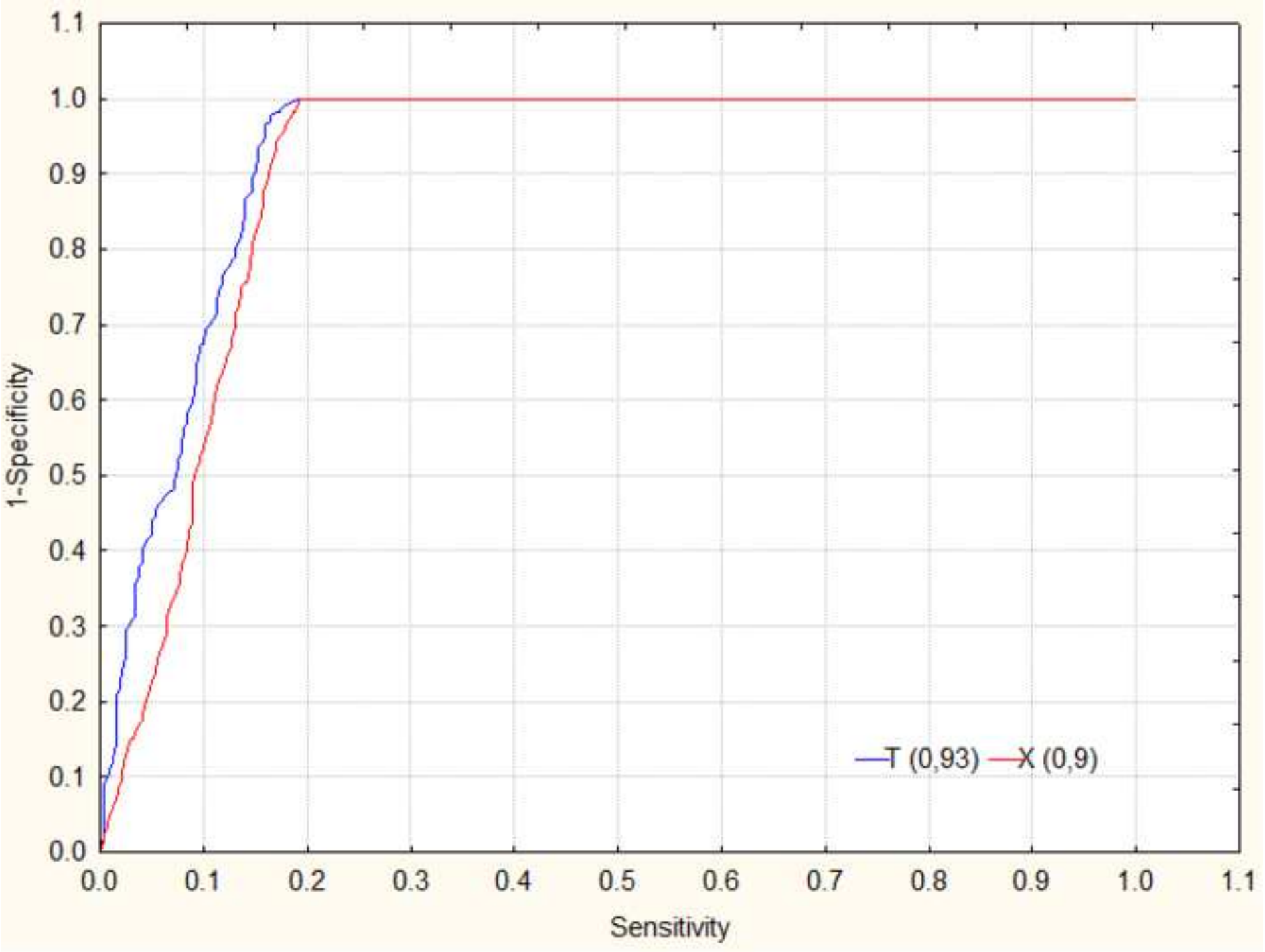
Figure(s)

Click here to download high resolution image



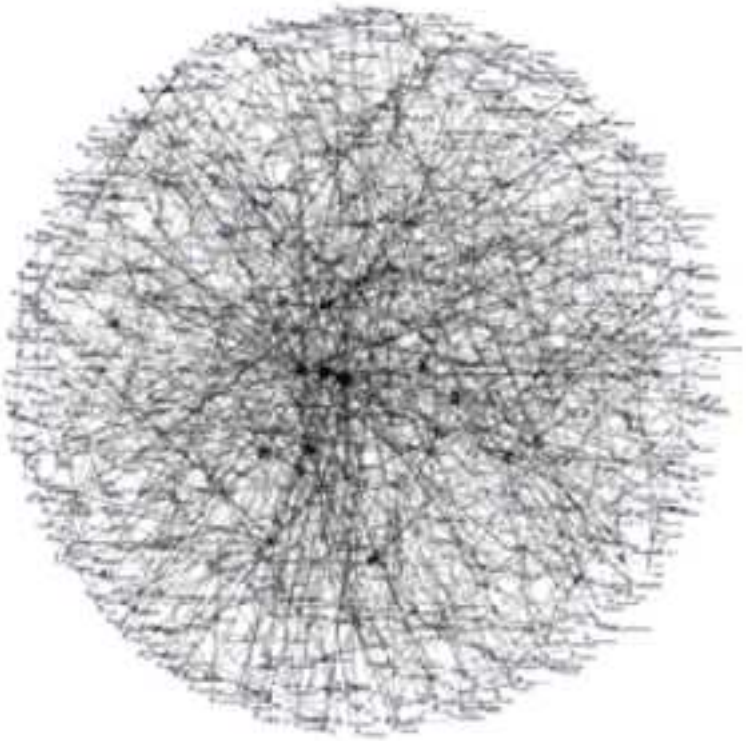
Figure(s)

Click here to download high resolution image

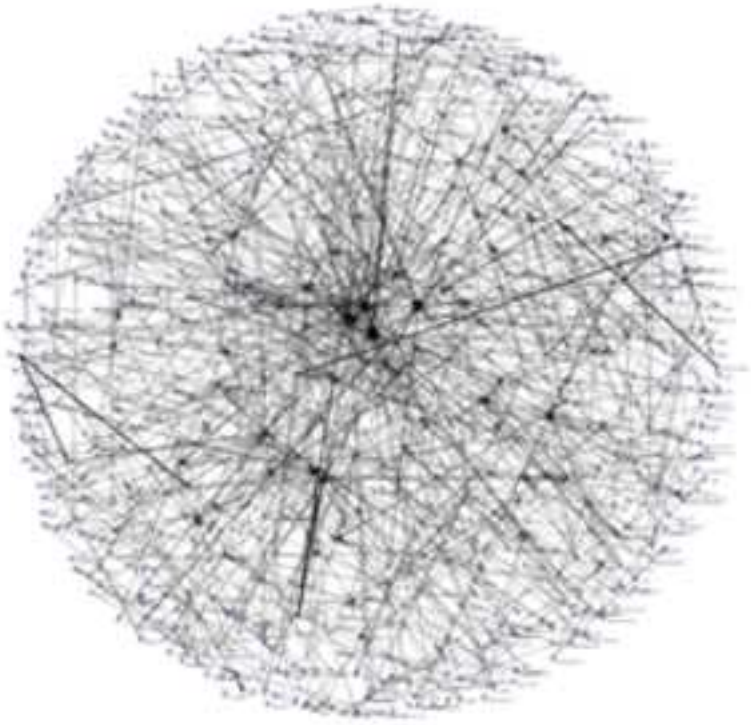


Figure(s)

Click here to download high resolution image



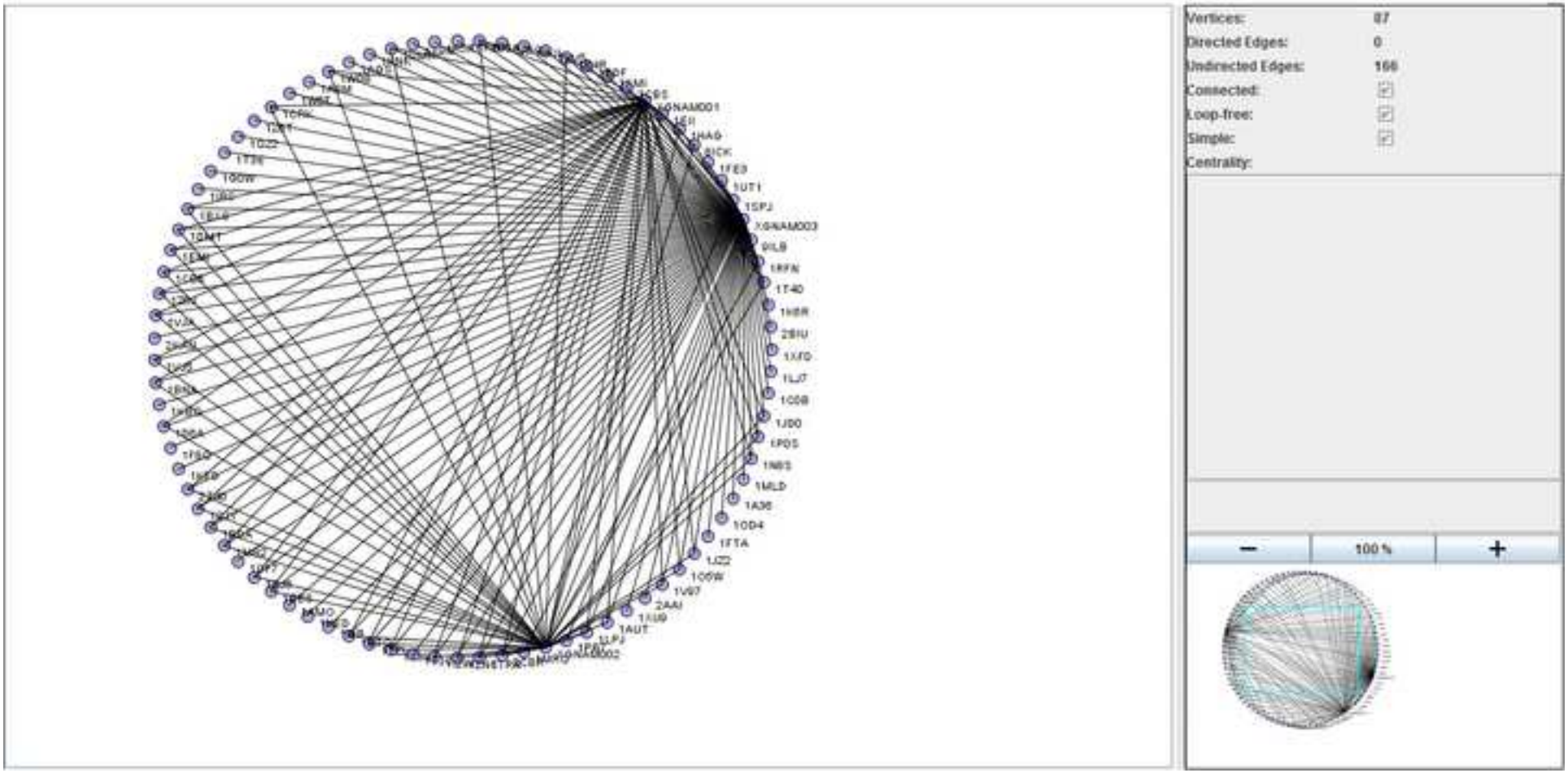
A



B

Figure(s)

Click here to download high resolution image



Journal of Theoretical Biology 293 (2012) 174–188

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

New Markov–Shannon Entropy models to assess connectivity quality in complex networks: From molecular to cellular pathway, Parasite–Host, Neural, Industry, and Legal–Social networks

Pablo Riera-Fernández [a], Cristian R. Munteanu [b], Manuel Escobar [c], Francisco Prado-Prado [c], Raquel Martín-Romalde [a], David Pereira [b], Karen Villalba [b], Aliuska Duardo-Sánchez [d], Humberto González-Díaz [a,*]

[a] Department of Microbiology & Parasitology, Faculty of Pharmacy, University of Santiago de Compostela (USC), 15782 Santiago de Compostela, Spain
[b] Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruña, Campus de Elviña, 15071 A Coruña, Spain
[c] Department of Organic Chemistry, Faculty of Pharmacy, USC, 15782 Santiago de Compostela, Spain
[d] Department of Special Public Law, Financial and Tributary Law Area, Faculty of Law, USC, 15782 Santiago de Compostela, Spain

ARTICLE INFO

Article history:
Received 22 July 2011
Received in revised form
9 October 2011
Accepted 14 October 2011
Available online 25 October 2011

Keywords:
Shannon Entropy
Markov Chains
Metabolic Pathways
Host–Parasite networks
Brain Cortex network

ABSTRACT

Graph and Complex Network theory is expanding its application to different levels of matter organization such as molecular, biological, technological, and social networks. A network is a set of items, usually called *nodes*, with connections between them, which are called *links* or *edges*. There are many different experimental and/or theoretical methods to assign node–node links depending on the type of network we want to construct. Unfortunately, the use of a method for experimental reevaluation of the entire network is very expensive in terms of time and resources; thus the development of cheaper theoretical methods is of major importance. In addition, different methods to link nodes in the same type of network are not totally accurate in such a way that they do not always coincide. In this sense, the development of computational methods useful to evaluate connectivity quality in complex networks (*a posteriori* of network assemble) is a goal of major interest. In this work, we report for the first time a new method to calculate numerical quality scores $S(L_{ij})$ for network links L_{ij} (connectivity) based on the Markov–Shannon Entropy indices of order k-th (θ_k) for network nodes. The algorithm may be summarized as follows: (i) first, the $\theta_k(j)$ values are calculated for all j-th nodes in a complex network already constructed; (ii) A Linear Discriminant Analysis (LDA) is used to seek a linear equation that discriminates connected or linked ($L_{ij}=1$) pairs of nodes experimentally confirmed from non-linked ones ($L_{ij}=0$); (iii) the new model is validated with external series of pairs of nodes; (iv) the equation obtained is used to re-evaluate the connectivity quality of the network, connecting/disconnecting nodes based on the quality scores calculated with the new connectivity function. This method was used to study different types of large networks. The linear models obtained produced the following results in terms of overall accuracy for network reconstruction: Metabolic networks (72.3%), Parasite–Host networks (93.3%), CoCoMac brain cortex co-activation network (89.6%), NW Spain fasciolosis spreading network (97.2%), Spanish financial law network (89.9%) and World trade network for Intelligent & Active Food Packaging (92.8%). In order to seek these models, we studied an average of 55,388 pairs of nodes in each model and a total of 332,326 pairs of nodes in all models. Finally, this method was used to solve a more complicated problem. A model was developed to score the connectivity quality in the Drug–Target network of US FDA approved drugs. In this last model the θ_k values were calculated for three types of molecular networks representing different levels of organization: drug molecular graphs (atom–atom bonds), protein residue networks (amino acid interactions), and drug–target network (compound–protein binding). The overall accuracy of this model was 76.3%. This work opens a new door to the computational reevaluation of network connectivity quality (collation) for complex systems in molecular, biomedical, technological, and legal–social sciences as well as in world trade and industry.

1. Introduction

Graph and Complex Network theory is expanding its application to different levels of matter organization such as molecular, biological,

* Corresponding author. Tel.: +34 981 167 000; fax: +34 981 167 160.
E-mail address: gonzalezdiazh@yahoo.es (H. González-Díaz).

0022-5193/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.10.016

technological, and social networks (Bornholdt and Schuster, 2003; Boccaletti et al., 2006; Dehmer and Emmert-Streib, 2009). A network is a set of items, usually called *nodes*, with connections between them, which are called *links* or *edges* (Newman, 2003). The nodes can be atoms, molecules, proteins, nucleic acids, drugs, cells, organisms, parasites, people, words, laws, computers, or any other part of a real system. The edges or links are relationships between the nodes such as chemical bonds, physical interactions, metabolic pathways, pharmacological action, law recurrence, or social ties.

There are many different experimental and/or theoretical methods to assign node–node links depending on the type of network we want to construct. Unfortunately, many of these methods are expensive in terms of time or resources. In addition, different methods to link nodes in the same type of network are not totally accurate in such a way that they do not always coincide. For instance, Modha and Singh, in their work 'Network architecture of the long-distance pathways in the macaque brain' (Modha and Singh, 2010) studied the information contained in the 'Collation of Connectivity data on the Macaque brain' (CoCoMac) neuroinformatic database in order to construct the most comprehensive long-distance network of the Macaque brain. This database contains 410 anatomical tracing studies, 10,681 connectivity relations and 16,712 mapping relations, and after collation of all connections, a final network of 383 brain regions and 6602 long-distance brain connections that travel through the brain's white matter were obtained. However, to construct this network, the authors had to solve problems related with the multiplicity of brain maps, divergent nomenclature, boundary uncertainty, different resolutions depending on the work studied. In this context, the development of fast and cheap computational methods in order to collate connectivity information becomes a goal of major importance.

One possible solution to this problem is the use of Quantitative Structure–Activity/Property Relationships (QSAR/QSPR) models, which have been traditionally studied in the field of chemoinformatics and are used to predict the biological activity of drugs (QSAR) or physicochemical properties of organic compounds (QSPR) using as input structural parameters of the system under study (Puzyn et al., 2010). In the case of global studies (properties of full system) these parameters are Topological Indices (TIs) derived from the graphical representation of the system (molecule, etc.). On the other hand, we can use node centralities or local TIs of a sub-graph if we want to predict a local property of part of the system (local chemical reactivity, biotransformation of a toxicophore group in a drug, etc.). Currently, the use of QSPR-like models in which the inputs are graph parameters is not limited to the study of molecules and has been extended to other complex systems (González-Díaz and Munteanu, 2010).

Specifically, Shannon entropy is one of the most useful parameters used as input in QSAR/QSPR studies to quantify structural information of molecular graphs (Dehmer et al., 2009). In all the above-mentioned cases, Shannon entropy parameters can be used to quantify structural information locally (nodes, edges, paths, clusters, etc.) and/or globally (full graph). In fact, we have used Markov Chain (MC) to calculate Shannon entropies locally or globally within a graph considering all possible branches at different topological distances. The information is quantified in terms of $\theta_k(j)$ values, which are called the Markov–Shannon entropy node centralities of order kth for all jth states (nodes) of a MC associated to the system. This MC is expressed by a Markov or Stochastic matrix ($\mathbf{\Pi}_1$) and represented by a graph of the studied system. The elements of $\mathbf{\Pi}_1$ are the probabilities ${}^1p_{ij}$ with which the ith and jth nodes connect each other (there is a physical or functional tie, link, or relationship) within a graph. Using Chapman–Kolmogorov equations it is straightforward to realize the way to calculate $\theta_k(j)$ values for all nodes in a graph. We can use these values directly or sum some of them to obtain total or local entropies (see Section 2). Our group has introduced the software called MARCH-INSIDE (Markovian Chemicals In Silico Design), which has become a very useful tool for QSAR/QSPR studies (Gonzalez-Diaz et al., 2010). This software can calculate 1D (sequence), 2D (connectivity in the plane) and 3D (connectivity in the space) MC parameters, including $\theta_k(j)$ values, for many molecular systems. MARCH-INSIDE is able to characterize small molecules (drugs, metabolites, organic compounds), biopolymers (gene sequence, proteins sequence or 3D structure, and RNA secondary structure) and artificial polymers but can perform a limited manage of other complex networks. It happens because MARCH-INSIDE can read, transform into Markov matrix, represent as graph, and calculate entropies for molecular formats (.mol or SMILE .txt files for drugs, .pdb for proteins, or .ct files for RNAs) but it is unable to upload formats of Complex Networks (.mat, .net, .dat, .gml, etc.).

In this work, we use for the first time QSPR-like models able to assess the quality of the connectivity of new complex networks assembled with information obtained from many sources not totally accurate. The idea is to seek a QSPR-like model that use as input the

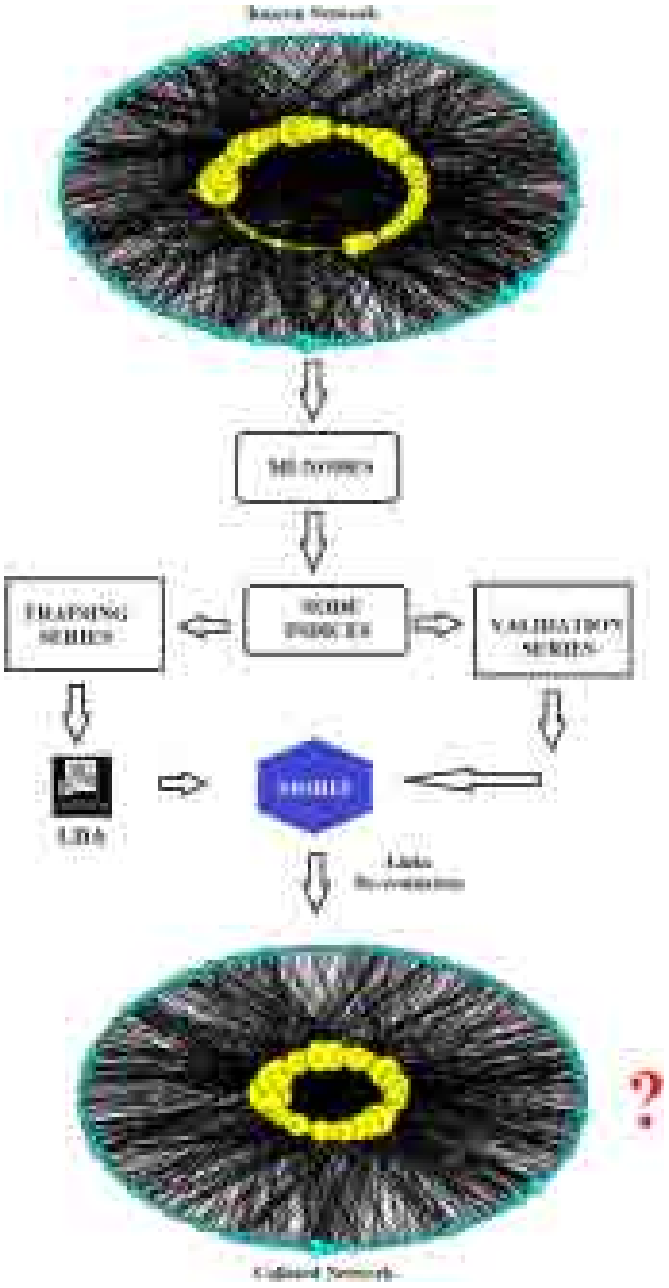


Fig. 1. General workflow used in this work.

$\theta_k(j)$ values for all possible pairs of nodes in a network to decide which pairs of nodes link each other and which ones do not. This class of model will allow us to computationally re-evaluate all the links in any complex network in such a way that we do not have to rely upon experimentation to confirm the existence or not of a link between all pairs of links. Using this model, we should experimentally confirm only those connections predicted by the model with low link score and/or simply remove them from the network depending on the cost/benefit ratio. As a consequence of this aim, we have re-programmed the MARCH-INSIDE application creating a new software able to manage complex networks. The new program is called MI-NODES (MARCH-INSIDE NOde DEScriptors) and is compatible with other software like Pajek or CentiBin, since it is able to read .mat, .net and .dat formats. A very interesting feature of MI-NODES is that it can process multiple networks within a file and calculate both MC global TIs and/or node centralities for all these networks. It is also able to export them in a single file in network-by-network and or node *vs.* node output formats. In order to illustrate the use of the new method we have carried out 7 experiments. In each experiment we report for the first time new QSPR models, which are useful to re-evaluate connectivity quality of different types of networks. Although very different systems were studied, the same workflow was used (see Fig. 1).

In the first experiment we studied the full metabolic pathway networks of four different organisms (bacteria, yeast, nematode, plant). In the second experiment, we studied different biological networks of parasite–host interactions (PHIs). The third experiment consisted of carrying out a study regarding connectivity quality in the CoCoMac cerebral cortex co-activation network (Modha and Singh, 2010). In the fourth experiment we studied a macroscopic landscape parasitism-spreading network for cattle fasciolosis in NW Spain. In the next experiment we illustrate the application of the method to a complex network for all the historic record of 1940–2004 of the entire Financial Law legislation system (legal–social network) in Spain. All these mentioned networks present one class of nodes or are bipartite. We also studied a 5th-partite network (with five classes of nodes) representing different relationships in the world trade of active & intelligent packaging for food industry. This additional network represents the relationships between companies, countries, trade mark products, products uses, and food types. Finally, we carried out an experiment to seek the first QSPR-like model for the US FDA Drug–Target network. This last experiment has methodological particularities because it involves different systems with graduated structural levels. We used as input θ_k values of drug molecular graph, protein structural networks, and drug–target network. With the advent of the age of complex systems science this study opens a door to a relatively less studied but very important field: the assessment of the connectivity quality in new complex networks. According a recent comprehensive review (Chou, 2011), to develop a useful predictor for a statistical system, the following things often need to be considered: (i) benchmark dataset construction or selection, (ii) formulation of statistical samples, (iii) operating algorithm (or engine), (iv) anticipated accuracy, and (v) web–server establishment. Below, let us elaborate how to deal with these procedures one by one.

2. Materials and methods

2.1. Datasets used

In the present study we have selected 7 types of networks, taking into account the data availability, the size of the networks studied (since we are interested in large complex networks) and the level of organization (from molecules to social sciences). Some of them have been used in previous studies and other are presented for the first time in this work.

2.1.1. Metabolic pathway networks

Metabolic network data was downloaded directly from Barabasi's group web (http://www.nd.edu/~networks/resources.htm) as gzipped ASCII file. In this file each number represents a substrate in the metabolic network of corresponding organism. Data-format is: From → To (directed link). The information studied was previously obtained by Jeong et al. from the 'intermediate metabolism and bioenergetics' portions of the WIT database and used in order to try to understand the large-scale organization of metabolic networks (Jeong et al., 2000). According to the authors, biochemical reactions described within a WIT database are composed of substrates and enzymes connected by directed links. For each reaction, educts and products were considered as nodes connected to the temporary educt–educt complexes and associated enzymes. Bidirectional reactions were considered separately. For a given organism with N substrates, E enzymes and R intermediate complexes the full stoichiometric interactions were compiled into an $(N+E+R)\,X\,(N+E+R)$ matrix, generated separately for each of the different organisms.

2.1.2. Parasite–Host complex networks

In order to construct the studied networks we have used two sources of information. The first of them is the Interaction Web Database (IWDB) (http://www.nceas.ucsb.edu/interactionweb/index.html), which contains datasets on species interactions from several communities in different parts of the world. In particular we have used the host–parasite dataset, composed of data belonging to studies about parasites (nematodes, acanthocephalans, cestodes, trematodes, monogeneans, leeches, copepods and branchiurans) and their hosts (fish) from 7 Canadian freshwater systems (Arai and Mudry, 1983; Arthur et al., 1976; Bangham, 1955; Chinniah and Threlfall, 1978; Dechtiar, 1972; Leong and Holmes, 1981). The second source of information is the Global Mammal Parasite Database (GMPD) (http://www.mammalparasites.org/), a compilation of records of parasites (helminths, protozoa, viruses, bacteria, arthropods and fungi) and their hosts (wild mammals) that have been documented in the published scientific literature (Nunn and Altizer, 2005). In this work we have used the information about ungulates (Order *Artiodactyla* and *Perissodactyla*), carnivores and primates. Based on the data obtained from the two databases, we constructed four bipartite networks (Parasite–Fish, Parasite–Ungulates, Parasite–Carnivores and Parasite–Primates) in which the first set of nodes is composed by parasites and the second by hosts, linked if the parasite interacts with the host.

2.1.3. Cerebral Cortex co-activation network

The version of the CoCoMac network used in this work consists of 383 hierarchically organized regions spanning cortex, thalamus, and basal ganglia; models the presence of 6602 directed long-distance connections (is three times larger than any previously derived brain network) and contains sub-networks corresponding to classic corticocortical, corticosubcortical, and subcortico-subcortical fiber systems (Modha and Singh 2010), see Fig. 2.

2.1.4. Complex network for fasciolosis spreading in NW Spain

The dataset reported by Mezo et al. (2008) in a previous work was used by our group to construct a network of farm-to-farm spreading of fasciolosis in cattle for Galicia (NW Spain) in other work (González-Díaz et al., 2010). In this work each farm was considered as a node of the network associated to a Boolean or connectivity matrix **C** with elements C_{ij} (links). As this is a

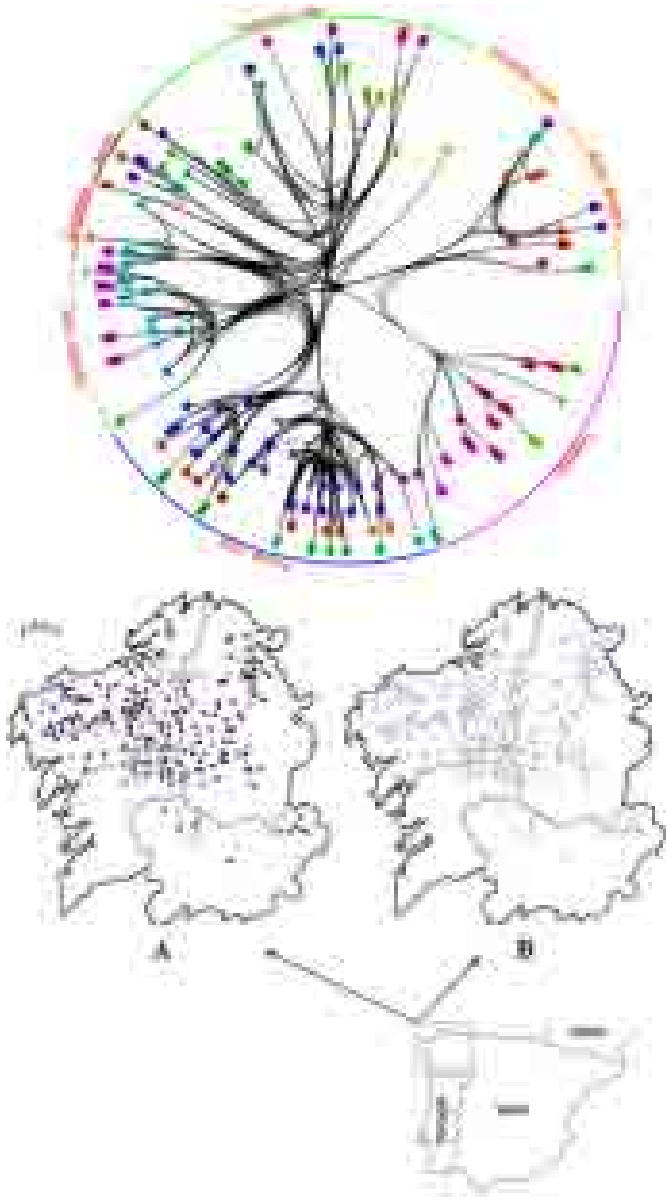


Fig. 2. Top: innermost core for the undirected version of CoCoMac network. The innermost core is a central sub-network that is far more tightly integrated than the overall network. Bottom: Geographical maps of Galicia (NW Spain) showing the location of the 275 sampled farms. **A**—Observed data for network **C**: the status of infection (empty circles: *F. hepatica* free and filled circles: *F. hepatica* infected) and the treatment administered on each farm are shown (blue: none; red: an anthelmintic effective against fluke mature stages and green: a fasciolicide effective against immature and mature stages). **B**—Observed data for network **L**: Distribution of farms according to the presence of *F. hepatica* infection (gray: uninfected; cyan: infected with a within-herd prevalence <25 % and pink: infected with a within-herd prevalence $\geq 25\%$).

symmetric condition, the existence of a connection *ij* implies the existence of the inverse connection *ji*. Loops, connections from *j*-th to the same *j*-th farm (representing self-infection of animals inside the same farm) were allowed. We place an arc (directed edge) connecting the *i*-th farm with the *j*-th farm if they meet the condition given in the Microsoft Excel command (see next equations) that is used to truncate the farm-to-farm distance function (see also Fig. 2). The connectivity of the network **C** depends on the input parameters: spatial coordinates (x_i, y_i) of the farm (f_i), the altitude of the place (h_i), and the strength of drug treatment ($Tr_i=0, 1, 2, 3$) used for pre-existent fasciolosis in this farm ($\mathrm{Tr}_i=0$ indicates that the disease was not detected). Consequently the matrix **C** quantifies the propensity $C_{ij}=1$ of the disease to spread between farms immediately after treatment. On the other hand, matrix **L** includes two criteria: the preexistence of a high propensity for disease spreading $C_{ij}=1$ and the experimental confirmation of a high Risk Ratio (RR_{ij}) of Prevalence After Treatment (PAT_j) of the disease. See the definition of these networks in mathematical terms using Excel functions:

$$L_{ij} = \mathrm{if}(\mathrm{AND}(C_{ij}=1, \mathrm{RR}_{ij}>1) \quad (1)$$

$$\mathrm{RR}_{ij} = (\mathrm{PAT}_i+1)/(\mathrm{PAT}_j+1) \quad (2)$$

$$C_{ij} = \mathrm{if}\ (\mathrm{OR}(d_{ij} > d_{\mathrm{cutoff}} * \frac{1}{m}\mathrm{Sum}(d_{ij}), d_{ij}=0), 0, 1) \quad (3)$$

$$d_{ij} = 0.5*(h_i+h_j)*Tr_i*Tr_j*\mathrm{SQR}((x_i-x_j)\hat{}2+(y_i-y_j)\hat{}2) \quad (4)$$

2.1.5. *Law co-recurrence network of the Spanish Financial-Legal system*

The studied network is built establishing connections between two laws or legal norms (nodes) if the time-lag is less than 1 for the same type of laws. Consequently, law–law links represent the co-recurrence of the Spanish Financial System along time to different norms depending on socio-economical conditions. The Cutoff function for the Spanish financial law recurrence network associated to the matrix **L** with elements L_{ij} is the following (see Fig. 3): L_{ij}=if (($D6–G$3) $>$$C$3, 0, if ($B6=G$1, 1, 0)). In this function $B6 and G$1 are Excel references to the column and row containing one-letter codes used to identify the type of financial law approved (node classes). The absolute Excel reference C3 points to the cell containing the time-lag cutoff value t_{off}. In addition, $D6 and G$3 are references to the column and row containing the values of the variable time (*yy*) equal to the year of approbation for a given law or norm. In Fig. 3 we illustrate the Excel sheet for the assembly of this Legal–Social network.

2.1.6. *Active & Intelligent Packaging for food industry*

Here we studied the world trade complex network for active & intelligent packaging in food industry (year 2011). This network interconnects product trade items of five classes (five node classes). The classes of nodes are: Product (PR), Company (CO), Country (CU), Food Type (FT), and product use identified as Packaging Type (PT). The network is 5th-partite in such a way that nodes of a given class are connected with nodes of other classes but nodes of the same class are never connected to each other. This network has been constructed, manually curated, and studied in a previous work. In this previous work the network was assembled using data obtained from several public resources previously compiled and reviewed in another paper (Pereira de Abreu et al., in press). The network created after these two works and used here contains a total of: 222 different products with registered trade mark (222 nodes of class PR), 60 different companies (CO), 15 countries (CU), 33 Food types (FTs) and 29 product types (PT). It makes a total of 359 nodes interconnected by 3868 links. The information encoded by a link depends on the classes of the two nodes interconnected. For instance, if one node belongs to the class CO and the other to the class PR, this link indicates that this company produces and commercializes this product.

2.1.7. *US FDA Drug–Target network*

In this case, the data used to built the drug–target network were obtained from the DrugBank database (http://www.drugbank.ca/), a bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical)

Fig. 3. Excel calculation sheet to obtain the matrix **L** of Spain Law Financial system **(A)** and MI-NODEs graphical interface **(B)**.

data with comprehensive drug target (i.e. sequence, structure, and pathway) information (Wishart et al., 2006, 2008; Wishart, 2010; Knox et al., 2011). In particular, 532 drugs approved by the Federal Food and Drug Administration (FDA) and 315 targets were studied.

2.2. Computational methods

2.2.1. Markov–Shannon entropy centralities for nodes

In information theory, entropy is a measure of the uncertainty associated with a random variable. The term by itself in this context usually refers to the Shannon entropy, which quantifies, in the sense of an expected value, the information contained in a message, usually in units such as bits. Equivalently, the Shannon entropy is a measure of the average information content one is missing when one does not know the value of the random variable. The concept was introduced by Claude E. Shannon in his 1948 paper "A Mathematical Theory of Communication" (Shannon 1948). In the present work, we construct the classical Markov matrix ($^1\mathbf{\Pi}$) for each network as follows. First, we downloaded from public resources the connectivity matrix **L** or obtain the data about the links between the nodes to assemble **L** (n by n matrix, where n is the number of vertices). Next, the Markov matrix **Π** is built. It contains the vertices probability (p_{ij}) based on **L**. The probability matrix is raised to the power k, resulting $(^1\mathbf{\Pi})^k$, and multiplied by the vector of the initial probabilities (^0p_j). The resulting vectors contain the absolute probabilities to reach the nodes moving throughout a walk of length k from node n_i (^kp_j) for each k and are the base for the entropy centrality (θ_k) calculation:

$$^kP = {}^0P(^1\Pi)^k = [{}^kp_1 + {}^kp_2 + \cdots + {}^kp_j] \tag{5}$$

$$\theta_k = -\sum {}^kp_j \log {}^kp_j \tag{6}$$

2.2.2. MI-NODES software for calculation of Markov–Shannon entropies

MI-NODES (MARCH-INSIDE NOde DEScriptors) is a GUI Python/wxPython application used for the calculation of a new class centralities/topological indices of nodes, sub-networks, or full networks. Actually, it should be considered as the

generalization of the software MARCH-INSIDE to manage any kind of complex networks (this program was originally designed to study drugs, proteins and nucleic acid structures). MI-NODES calculates new types of node Centralities ${}^{k}C_c(j)$ based on Markov normalized node probabilities without removing each node previously to perform calculations. It also calculates Markov generalizations of different topological indices ${}^{k}TI_c(G)$ of class c and power k for the graph G. The tool is both Pajek and CentiBin compatible because it reads networks in the following formats: .net, .dat and .mat. We depict MI-NODES interface and an example of one of the steps in the assembly of a .mat file in Fig. 3. MI-NODES can calculate the following types of Markov node centralities and topological indices of order k-th: Shannon Entropy, Spectral Moments, Harary numbers, Wiener indices, Gutman topological indices, Schultz topological indices, Broto indices, Balaban indices, Kier–Hall connectivity indices, Randić connectivity indices, Galvez indices and Leverage indices.

2.2.3. Dataset used to construct the model

The first step to obtain the dataset for each model was to calculate the θ_k values for the linked nodes of the networks of each type from $K=0$ to 5 using MI-NODES. These are the positive cases. In the second step, the same was done but with the negative cases (nodes that are no linked in the observed network). These negative cases were chosen randomly by MI-NODES (we did not take into account 100% of them because their number is much higher than the number of positive cases and this difference can have negative effects on the statistical analysis). Finally, 75% of the data (chosen randomly) was used for training the model and the remaining 25% for cross-validation.

2.2.4. Linear Discriminant Analysis (LDA) models

Once the values of the Markov–Shannon entropies were obtained, we carried out a Linear Discriminant Analysis (LDA) by means of the STATISTICA software (StatSoft.Inc. et al., 2002). LDA is possibly the most common technique used in QSPR/QSAR studies with TIs of molecular graphs, protein and RNA structure networks, and bio-molecular complex networks. Let $S(L_{ij})$ be the output variable of a model used to score the quality of the connection between two nodes i-th and j-th ($L_{ij}=1$). We can use LDA to seek a linear equation with coefficients a_i, a_j, a_{ij} and a_0. These are the coefficients of the TIs used as input (in this case local node centralities) in the QSAR/QSPR model and the independent term. The terms a_{ik} and a_{jk} refer to all nodes that lie within the k-th neighborhood (placed at least at topological distance $d=k$) of i-th or j-th single nodes, respectively. The term a_{ijk} refers to the differences between the neighborhoods of a pair of nodes, which may be connected or not. We can use different statistical parameters to evaluate the statistical significance and validate the goodness-of-fit of LDA equation: n=number of cases, χ^2=Chi-square, p=the error level, as well as the Accuracy, Specificity, and Sensitivity of both train and external validation series (Hill and Lewicki, 2006). Generally, we write the linear LDA equation with the parameters mentioned above in the following form (example using the entropy values as TIs), see also Fig. 1:

$$S(L_{ij}) = \sum_{k=0}^{5} a_{ik}\theta_k(i) + \sum_{k=0}^{5} a_{jk}\theta_k(j) + \sum_{k=0}^{5} a_{ijk}[\theta_k(i)-\theta_k(j)] + a_0 \quad (7)$$

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test (Chou and Zhang, 1995). However, of the three test methods, the jackknife test is deemed the most objective (Chou and Shen, 2008). The reasons are as follows. (i) For the independent dataset test, although all the proteins used to test the predictor are outside the training dataset used to train it so as to exclude the 'memory' effect or bias, the way of how to select the independent proteins to test the predictor could be quite arbitrary unless the number of independent proteins is sufficiently large. This kind of arbitrariness might result in completely different conclusions. For instance, a predictor achieving a higher success rate than the other predictor for a given independent testing dataset might fail to keep so when tested by another independent testing dataset (Chou and Zhang, 1995). (ii) For the subsampling test, the concrete procedure usually used in literatures is the 5-fold, 7-fold or 10-fold cross-validation. The problem with this kind of subsampling test is that the number of possible selections in dividing a benchmark dataset is an astronomical figure even for a very simple dataset, as elucidated by Chou and Shen (2008) and demonstrated by Eqs. 28–30 in Chou (2011). Therefore, in any actual subsampling cross-validation tests, only an extremely small fraction of the possible selections are taken into account. Since different selections will always lead to different results even for a same benchmark dataset and a same predictor, the subsampling test cannot avoid the arbitrariness either. A test method unable to yield a unique outcome cannot be deemed as a good one. (iii) In the jackknife test, all the proteins in the benchmark dataset will be singled out one-by-one and tested by the predictor trained by the remaining protein samples. During the process of jackknifing, both the training dataset and testing dataset are actually open, and each protein sample will be in turn moved between the two. The jackknife test can exclude the 'memory' effect. Also, the arbitrariness problem as mentioned above for the independent dataset test and subsampling test can be avoided because the outcome obtained by the jackknife cross-validation is always unique for a given benchmark dataset. Accordingly, the jackknife test has been increasingly and widely used by those investigators who have strong math background to examine the quality of various predictors (see, e.g., (Chen et al., 2009; Chou and Shen, 2010; Chou et al., 2011; Gu et al., 2010; Lin et al., 2011; Mohabatkar 2010; Wang et al., 2011; Xiao et al., 2011; Zakeri et al., 2011; Zeng et al., 2009; Zhang et al., 2011)). However, to reduce the computational time, we adopted the independent testing dataset cross-validation in this study as done by many investigators with support vector machine (SVM) as the prediction engine.

2.2.5. Graphical representation and description of the networks

Using graphical/diagrammatic approaches to study complicated systems can provide an intuitive picture or useful insights to help in analyzing complicated mechanisms in these systems, as demonstrated by many previous studies on a series of important biological topics, such as enzyme-catalyzed reactions (Andraos, 2008; Chou, 1989; Chou and Forsen, 1980; Zhou and Deng, 1984), protein folding kinetics and folding rates (Chou, 1990), inhibition of HIV-1 reverse transcriptase (Althaus et al., 1993a; Althaus et al., 1993), inhibition kinetics of processive nucleic acid polymerases and nucleases (Chou et al., 1994), drug metabolism systems (Chou 2010), analysis of DNA sequence (Xie and Mo, 2011), and protein sequence evolution (Wu et al., 2010). Recently, the wenxiang diagrams (Chou et al., 1997) were also used to investigate protein–protein interactions (Zhou, 2011a, 2011b).

In order to characterize the studied systems, we have represented graphically and calculated some parameters of both the observed and reconstructed networks. The information to construct them was obtained from the connectivity matrices in the case of the observed networks and from the output of the LDA models in the case of the reconstructed networks. CentiBin (http://centibin.ipk-gatersleben.de/index.php) (v.1.4.3)

(Junker et al., 2006) was used to prepare the networks using the command: Utils → 'prepare for centralities-undirected'. This command performs the following steps: removal of existing loops, removal of parallel edges, reduction to the giant component and transformation into an undirected network. Once the networks were prepared, it was possible to calculate the graph diameter, Wiener index and average distance. The density (taking into account if the graph is unipartite or bipartite), average degree and Randić index were also calculated, but using Pajek (v.1.26) (http://vlado.fmf.uni-lj.si/pub/networks/pajek/) (Batagelj and Mrvar 1998; De Nooy et al., 2005). This program was used to represent graphically the prepared versions of the observed and reconstructed networks.

3. Results and discussion

As can be seen in the next sub-sections, the models developed using structural information encoded by θ_k values offers good results in terms of overall accuracy (ranging from 72.3% to 97.2%) for all the studied networks. However, there are important differences between the best and the worst result (24.9%). This fact could suggest that the capacity to encode information by θ_k is strongly influenced by the structure of the network (similar values for θ_k in positive and negative cases could result in a low discrimination power) or that the linear model is less suitable for some types of networks. Anyway, more studies are needed to measure the influence of different factors on the performance of the models based on θ_k and other TIs. The use of each model developed in this section is restricted to networks with the same features than the networks used to develop the models. For example, the metabolic pathway model (constructed taking into account four model organisms) could be used to evaluate the equivalent metabolic networks of other organisms. An interesting possibility for future research would be to study the same networks with other TIs to see if the results improve (for example models with an accuracy of 70%) or to seek models in which various TIs are combined (in this case each type of TI could encode a different type of structural information).

3.1. Model 1: metabolic pathway networks

Study of metabolic networks is of great interest in biology because many applications are directly built on the use of cellular metabolism. Biotechnologists modify the cells and use them as cellular factories to produce antibiotics, industrial enzymes, antibodies, etc. In biomedicine, it is possible to cure metabolic diseases through a better understanding of the metabolic mechanisms, and to control infections by making use of the metabolic differences between human beings and pathogens (Rosa da Silva et al., 2008). For example, the network topology-based approach has been used to uncover shared mechanisms in the study of disease comorbidity (Lee et al., 2008). In a cell or microorganism, metabolic pathways are seamlessly integrated through a complex network of cellular constituents and reactions. However, despite the key role of these networks in sustaining cellular functions, their large-scale structure is not well known. Jeong et al. (2000) showed that, despite significant variation in their individual constituents and pathways, these metabolic networks have the same topological scaling properties and show striking similarities to the inherent organization of complex non-biological systems. An interesting question to answer is: given a regulatory pathway system consisting of a set of proteins, can we predict which pathway class it belongs to? In this sense, Huang et al. (2011) developed a computational method for the classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property. In another work carried out by the same author (Huang et al., 2010), a computational method was developed for the analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. Many pathways are not totally confirmed experimentally but have been computationally deduced using protein or gene alignment techniques. The idea follows more or less the following scheme: similar proteome → similar enzymes → similar metabolome. On the other hand, the experimental determination of the full metabolome including each metabolite and metabolite bio-transformation pathways is not always an easy task. All this aspects determine the necessity of alignment-free techniques to assess network connectivity quality in existing models of metabolic pathway networks. Here we developed a model to re-evaluate connectivity using as inputs the θ_k values for nodes in already-known metabolic networks. For this analysis we have used metabolic networks of four model organisms belonging to different domains of the tree of life. These organisms are: *Escherichia coli* (EC), *Saccharomyces cerevisiae* (SC), *Caenorhabditis elegans* (CE), and *Oryza sativa* (OS). *E. coli* is a gram negative bacterium that is commonly found in the lower intestine of warm-blooded organisms. Most EC strains are harmless, but some serotypes can cause serious food poisoning in humans. From the point of view of research, it is one of the best studied bacteria, especially in the areas of genetics, biochemistry and metabolism, and many researchers have based their studies on existing metabolic networks of this prokaryotic model (Baldazzi et al., 2010; Costa et al., 2010; Gerlee et al., 2009; Fowler et al., 2009; Konig et al., 2006; Imielinski et al., 2006; Shi et al., 1999; Lin et al., 2005; Ghim et al., 2005; Schmid et al., 2004; Light and Kraulis, 2004; Burgard and Maranas, 2001; Edwards and Palsson, 2000). *S. cerevisiae* (a species of yeast) is a fungus with industrial importance used as a model for understanding and engineering eukaryotic cell function. In fact, it was the first eukaryotic genome that was fully sequenced, annotated, and made available publicly (Goffeau, 1997). *C. elegans* is a free-living nematode that has become a popular model for genetic and molecular research, since it is easy to maintain and has a very fast life-cycle (Burglin et al., 1998). It was the first multi-cellular organism to have its genome completely sequenced (Consortium TCeS et al., 1998). In the field of parasitology, comparison between CE and other parasitic nematodes is an interesting method for studying the function and regulation of some parasite genes (Bird and Opperman, 1998). Other interesting feature is that CE is sensitive to the majority of anti-helmintic drugs that are used against parasitic worm infections of humans and livestock. This has provided the opportunity to use molecular genetic techniques in the worm for mode of action studies (Holden-Dye and Walker, 2007). Finally, *O. sativa*, commonly known as rice, is a plant of the family *Poaceae* with great economic importance for the human being. Over recent years, it has gained importance as a model organism for genetic and molecular studies. This is due to its relatively small genome of 420 Mb, whose full sequence was released as early as 2002. The many tools and experimental approaches now available for rice have made it the most widely studied model for cereals (Muller and Grossniklaus, 2010).

The best model found was

$$S(L_{ij}) = 159.16\theta_3(e_i) - 120.70\theta_1(p_j) - 95.42[\theta_5(e_i) - \theta_5(p_j)] - 0.26$$

$$n = 74{,}999 \quad \chi^2 = 26{,}093 \quad p < 0.001 \tag{8}$$

In this equation, $S(L_{ij})$ is a real-valued output variable that scores the propensity of the ith input or educt (e_i) (reactant or substrate) to undergo a metabolic transformation into the product (p_j). The parameter θ_1 quantifies the information related to

Table 1
Training and Cross-validation results for all models developed in this work.

QSPR	Training series				Model	Cross-validation series			
model	NL	L		%	Parameters	%		NL	L
1	**Metabolic pathway networks**								
	46,029	18,490	**NL**	71.3	Specificity	71.5	**NL**	15,384	6,123
	2,295	8,185	**L**	78.1	Sensitivity	77.8	**L**	775	2,719
			Total	72.3	Accuracy	72.4	**Total**		
2	**Parasite-host networks**								
	42,576	2,052	**NL**	95.4	Specificity	95.6	**NL**	14,144	652
	1,275	3,315	**L**	72.2	Sensitivity	71.3	**L**	436	1,085
			Total	93.2	Accuracy	93.3	**Total**		
3	**Cerebral Cortex co-activation network**								
	31,886	2,698	**NL**	92.2	Specificity	92.5	**NL**	10,637	867
	1,425	3,527	**L**	71.2	Sensitivity	70.4	**L**	488	1,162
			Total	89.6	Accuracy	89.7	**Total**		
4	**NW Spain Fasciolosis Landscape-Spreading network**								
	18,153	149	**NL**	99.2	Specificity	99.1	**NL**	6,068	58
	405	964	**L**	70.4	Sensitivity	74.2	**L**	116	334
			Total	97.2	Accuracy	97.4	**Total**		
5	**Legal–social network of the Spanish financial law system**								
	15,564	3,401	**NL**	82.1	Specificity	81.6	**NL**	5,172	1,169
	0	14,986	**L**	100.0	Sensitivity	100.0	**L**	0	5,014
			Total	90.0	Accuracy	89.7	**Total**		
6	**World Trade Intelligent-active food packaging network**								
	27,387	1,623	**NL**	94.4	Specificity	94.4	**NL**	9,128	542
	692	2,209	**L**	76.1	Sensitivity	77.5	**L**	218	749
			Total	92.7	Accuracy	92.9	**Total**		
7	**US FDA Drug-target network**								
	3,206	981	**NL**	76.6	Specificity	77.8	**NL**	1,079	308
	189	532	**L**	73.8	Sensitivity	70.4	**L**	71	169
			Total	76.2	Accuracy	76.7	**Total**		

Rows: Observed classifications; Columns: Predicted classifications; L: Linked; NL: Not linked.

the position of the input or reactant metabolite and their direct neighbors ($k=1$) in the metabolic network. The parameter θ_5 quantifies the information related to middle-long range subsequent metabolic transformations of all the neighbors of the product metabolite ($k=5$) in the metabolic network. As we can see in the previous equation the $\chi^2=26{,}093$ statistic corresponds to a p-level <0.001, which indicates a significant discrimination between known metabolic reactions and those metabolite transformations which are not experimentally observed. The model presents very good values of Accuracy, Sensitivity, and Specificity for the recognition of links both in training and external validation series see Table 1.

In Table 2 we carry out a graphical and numerical comparison of the giant components of the metabolic networks already known vs. those obtained after re-evaluating link quality with our model.

3.2. Model 2: Parasite–Host networks

Due to the importance for the human and animal health and therefore for the economy, much attention has been focused on parasite–host interactions (PHIs). The study of these interactions can help us to understand the role of phylogenetic and ecological factors on the parasite–host specificity (Desdevises et al., 2002; Detwiler and Janovy, 2008; Poulin et al., 2011) and to know how parasites affect the ecosystem functioning (Hatcher et al., 2006; Price et al., 1986; Anderson and May, 1979). In this sense, network theory is a useful tool for analyzing this type of interactions (Poulin, 2010). However, the high experimental difficulty inherent to the *in situ* accurate determination of PHIs makes the possibility of curate PHIs networks using a computational model very interesting. In this work, we used θ_k to seek a QSPR-like model able to score the quality of PHIs in known networks. The best model found was

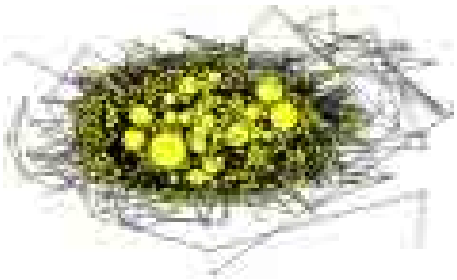
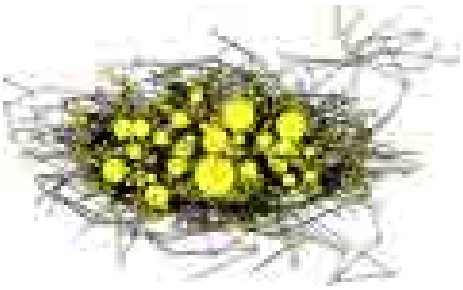
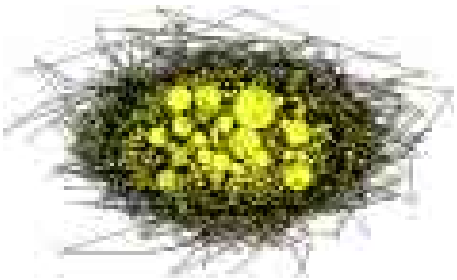
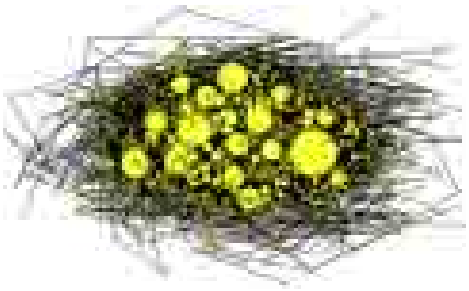
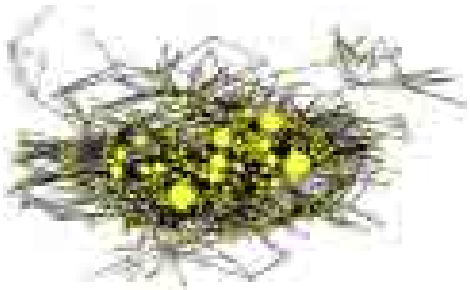
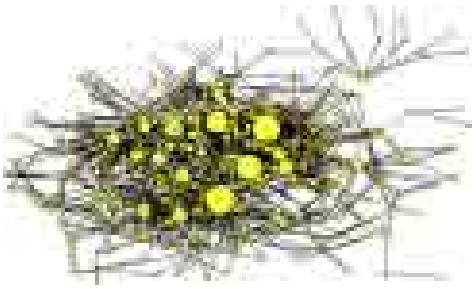
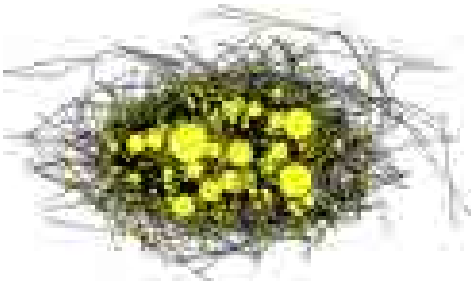
$$S(L_{ij}) = -82.62[\theta_5(p_i)-\theta_5(h_j)]-5.52$$
$$n=49{,}218 \quad \chi^2 = 21{,}728 \quad p < 0.001 \tag{9}$$

In this equation, $S(L_{ij})$ is a real-valued output variable that scores the propensity of the ith parasite specie (p_i) to infect a given host specie (h_j). The Chi-square statistic (χ^2) presents a low p-level <0.001, which indicates a significant discrimination between well-established host–parasite relationships and not confirmed parasitism. The model presents good values of Accuracy, Sensitivity, and Specificity for the recognition of parasite–host relationships (links) both in training and external validation series (see Table 1). Consequently, with this simple linear model we could re-evaluate connectivity quality in the already-known PHIs networks in a fast and non-expensive way (without to experimentally resample all PHIs in the correspondent ecological niche). The giant components of the observed and reconstructed networks are described numerically and graphically in Table 3.

3.3. Model 3: Cerebral Cortex co-activation network

Connectivity is the key to understanding distributed and cooperative brain functions. Detailed and comprehensive data on large-scale connectivity between primate brain areas have been collated systematically from published reports of experimental tracing studies (Kotter, 2004). Databasing the brain's anatomical connectivity as delivered by tracing studies is of particular importance as these data characterize fundamental

Table 2
Comparison of observed *vs.* re-constructed metabolic pathway networks (giant components).

Observed network	Network descriptors[a]			Reconstructed network
Caenorhabditis elegans				
	1,173	n	1,037	
	2,842	m	2,214	
	4.85	Ad	4.27	
	0.00413	den	0.00412	
	479.26	R	371.21	
	6,304,312	W	4,868,792	
	14	D	14	
	4.59	AD	4.53	
Escherichia coli				
	2,268	n	1,991	
	5,620	m	4,393	
	4.96	Ad	4.41	
	0.00219	den	0.00222	
	871.81	R	646.79	
	22,914,068	W	17,450,524	
	12	D	12	
	4.46	AD	4.40	
Oryza sativa				
	658	n	566	
	1,498	m	1,185	
	4.55	Ad	4.19	
	0.00693	den	0.00741	
	282.24	R	221.01	
	2,061,882	W	1,494,724	
	14	D	12	
	4.77	AD	4.67	
Sacharomices cerevisae				
	1511	n	1295	
	3807	m	2928	
	5.04	Ad	4.52	
	0.00334	den	0.00349	
	600.49	R	433.95	
	10,332,580	W	7,455,256	
	14	D	14	
	4.53	AD	4.45	

[b]The size of each node is proportional to its normalized degree.

[a] Network descriptors: Total number of connected nodes (n), number of edges (m), average degree (Ad), density (den), Randić connectivity index (R), Wiener index (W), diameter (D) and average distance (AD).

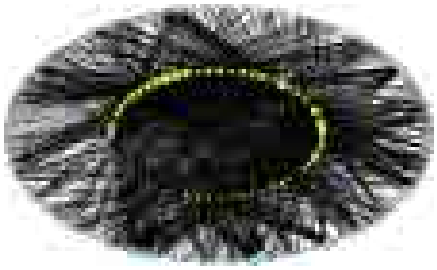
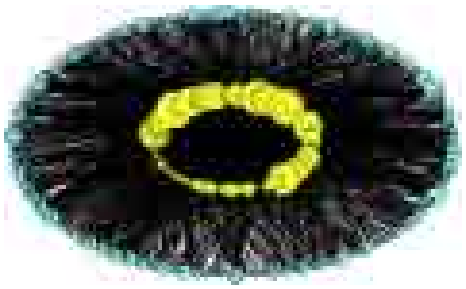
structural constraints of the complex and poorly understood functional interactions between the components of real neural systems. The eventual impact and success of connectivity databases, however, will require the resolution of several methodological problems that currently limit their use. These problems comprise four main points: (i) objective representation of coordinate-free, parcellation-based data, (ii) assessment of the reliability and precision of individual data, especially in the presence of contradictory reports, (iii) data mining and integration of large sets of partially redundant and contradictory data, and (iv) automatic and reproducible transformation of data between incongruent brain maps (Stephan et al., 2001). In order to address points ii and iv, we have developed a specific model for the 'collation of connectivity data on the macaque brain' (CoCoMac) database (http://www.cocomac.org). The best model found was

$$S(L_{ij}) = 70.56\theta_1(i) + 74.51\theta_5(j) - 1.75$$

$$n = 39{,}536 \quad \chi^2 = 22{,}249 \quad p < 0.001 \tag{10}$$

In this equation, $S(L_{ij})$ is a real-valued output variable that scores the propensity of the ith cerebral cortex region to undergo co-activation with the jth region in the CoCoMac network.

Table 3
Comparison of observed *vs.* re-constructed parasite-host interaction networks (giant components).

Observed network	Network descriptors [a]		Re-constructed network
Parasites–fish			
	298	n	271
	239	np	233
	59	nh	38
	912	m	788
	6.12	Ad	5.82
	0.0647	den	0.0890
	95.77	R	80.91
	298,626	W	246,480
	7	D	6
	3.37	AD	3.37
Parasites–Ungulates			
	793	n	675
	701	np	645
	92	nh	30
	1863	m	1393
	4.70	Ad	4.13
	0.0289	den	0.0720
	191.02	R	125.79
	2,534,140	W	1,763,848
	10	D	6
	4.03	AD	3.88
Parasites–carnivores			
	619	n	537
	537	np	505
	82	nh	32
	1343	m	1074
	4.34	Ad	4.00
	0.0305	den	0.0665
	159.16	R	115.30
	1,587,048	W	1,179,176
	9	D	8
	4.15	AD	4.10
Parasites–primates			
	913	n	612
	757	np	582
	156	nh	30
	1,993	m	1145
	4.37	Ad	3.74
	0.0169	den	0.0656
	252.66	R	118.58
	3,609,008	W	1,450,756
	10	D	8
	4.33	AD	3.88

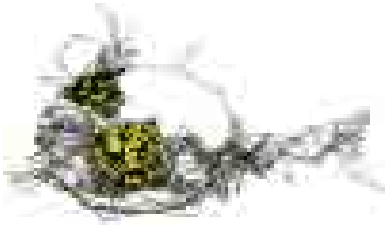
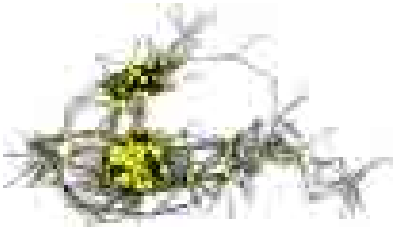
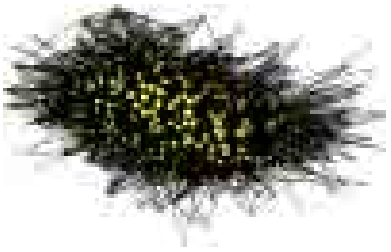
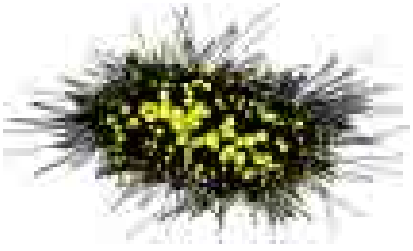
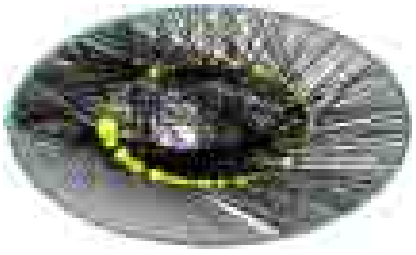
[a] Network descriptors: Total number of connected nodes (n), number of connected parasites (np), number of connected hosts (nh), number of edges (m), average degree (Ad), density (den), Randić connectivity index (R), Wiener index (W), diameter (D) and average distance (AD). [b]The size of each node is proportional to its normalized degree. [c] Outer nodes: Parasites; Inner nodes: Hosts.

The parameter $\theta_1(i)$ quantifies the information related to the position of the ith region and their direct neighbors ($k=1$) in the network. The parameter $\theta_5(j)$ quantifies the information related to middle-long range co-activation of different brain areas ($k=5$) in the cerebral cortex. As in the previous equation the $\chi^2=22{,}249$ statistics corresponds to a p-level <0.001, which indicates a significant discrimination between co-activated regions and not co-activated ones. The model presents very good values of Accuracy, Sensitivity, and Specificity (see Table 1). The giant components of the observed and reconstructed networks are described numerically and graphically in Table 4.

3.4. Model 4: Fasciolosis spreading network (NW Spain)

Fasciolosis is a parasitic infection caused by *Fasciola hepatica* (liver fluke) that has become an important cause of lost productivity in livestock worldwide. Considered a secondary zoonotic disease until the mid-1990s, human fasciolosis is at present

Table 4
Comparison of observed *vs.* re-constructed networks (giant components).

Observed network	Network descriptors [a]		Reconstructed network
Fasciolosis Landscape-Spreading Network for NW Spain			
	259	n	252
	1,798	m	1,278
	13.88	Ad	10.14
	0.0538	den	0.0404
	114.47	R	91.50
	273,596	W	260,884
	11	D	10
	4.09	AD	4.12
			
Cerebral Cortex co-activation Network (CoCoMac)			
	360	n	351
	5208	m	3,720
	28.93	Ad	21.20
	0.0806	den	0.0606
	145.98	R	115.90
	292,892	W	284,238
	5	D	4
	2.27	AD	2.31
			
Financial Law Network			
	223	n	223
	16,611	m	16,611
	148.98	Ad	148.98
	0.671	den	0.671
	111.12	R	111.12
	65,790	W	65,790
	2	D	2
	1.33	AD	1.33
			
US FDA Drug–Target Network			
	638	n	448
	404	nd	355
	234	nt	93
	802	m	527
	2.51	Ad	2.35
	0.00848	den	0.0160
	213.33	R	133.41
	2,702,728	W	1,449,898
	17	D	18
	6.65	AD	7.24
			

[a] Network descriptors: Total number of connected nodes (n), number of connected drugs (nd), number of connected targets (nt), number of edges (m), average degree (Ad), density (den), Randić connectivity index (R), Wiener index (W), diameter (D) and average distance (AD). [b] The size of each node is proportional to its normalized degree. [c] In the Drug–Target network the outer nodes represent drugs and the inner nodes targets.

emerging or re-emerging in many countries, including increases of prevalence and intensity and geographical expansion. In fact, research in recent years has justified the inclusion of fasciolosis in the list of important human parasitic diseases. At present, fasciolosis is the vector-borne disease presenting the widest latitudinal, longitudinal and altitudinal distribution known. In addition, it presents a range of epidemiological characteristics related to a wide diversity of environments (Mas-Coma, 2005). In this sense, the study of geographical spreading of fasciolosis becomes a subject of great interest. In fact, in a recent work we have constructed a network to study the landscape spreading of fasciolosis in Galicia (NW Spain) (González-Díaz et al., 2010). However, we do not have quantitative criteria on the quality of the network connectivity, and re-sampling of all data to re-evaluate this connectivity in a field study is a hard and expensive task in terms of time and resources. This situation has prompted us to seek a model in order to assess the quality of the network previously assembled. The best QSPR model found was

$$S(L_{ij}) = -20.23\theta_1(f_i) + 165.13\theta_4(f_j) - 0.82$$

$$n = 19{,}671 \quad \chi^2 = 16{,}058 \quad p < 0.001 \tag{11}$$

The entropy values $\theta_k(f_i)$ and $\theta_k(f_j)$ used in this equation quantify information about the connectivity patterns between farms in the network **C**. As can be seen in the equations described in Section 2, the connectivity of **C** depends on the spatial coordinates (x_i, y_i) of the farm (f_i), the altitude of the place (h_i), and the anti-parasite drug treatment (Tr_j) used to prevent Fasciolosis in this farm. Consequently the matrix **C** quantifies the *a priori* propensity $C_{ij}=1$ of this disease to spread between farms immediately after treatment depending on geographical conditions. On the other hand, matrix **L** includes both criteria: (i) the preexistence of a high propensity for disease spreading $C_{ij}=1$ and (ii) the experimental confirmation $L_{ij}=1$ of a high Risk

Ratio (RR_{ij}) of Prevalence After Treatment (PAT_j) for this disease in farms (see Fig. 2). The QSPR equation developed here was obtained by studying **L** and the model presents good values of Accuracy, Sensitivity, and Specificity (see Table 1). Both observed and reconstructed networks (giant components) are described and represented graphically in Table 4.

3.5. Model 5: Legal–Social network for the Spanish Financial Law system

The use of network analysis methods in social sciences began in 1930 and today are widely used (Wasserman and Faust, 1999). However, the application of these methods in legal studies is still at the beginning (Fowler and Jeon, 2008; Duardo-Sánchez, 2010; Duardo-Sánchez, 2011). Network tools may illustrate the interrelation between the different law types and help to understand law consequences in society and its effectiveness or not. We have used the list of the financial laws to construct the network described. The best model found was

$$S(L_{ij}) = 650.88[\theta_1({}^cL_{ti}) - \theta_1({}^cL_{ti+1})] + 0.12$$
$$n = 33{,}951 \quad \chi^2 = 32{,}942 \quad p < 0.001 \qquad (12)$$

where $\theta_k(L_{ti})$ and $\theta_k(L_{ti+1})$ are the entropy parameters that quantify information about the Legal norms (Laws) of type L introduced in the Spanish legal system at time t_i and t_{i+1} with respect to the previous or successive kth norms approved. The model behaves like a time series embedded within a complex network. This is because it predicts the recurrence of the Spanish law system to a financial norm of class c when socio-economical conditions change at time t_{i+1} given that have been used a known class of norm in the past at time t_i. The model correctly reconstructed the network of the historic record for the Spanish financial system with high Accuracy, Specificity, and Sensitivity (Table 1). The graphical representation and calculus of network descriptors of the observed and reconstructed networks (giant components) can be seen in Table 4.

3.6. Model 6: World Trade Network of Active & Intelligent Packaging for Food Industry

Traditionally, the basic functions of packaging have been classified into 4 categories: protection, communication, convenience, and containment. The package is used to protect the product against the deteriorative effects of the external environment, communicate with the consumer as a marketing tool, provide the consumer with greater ease of use and time-saving convenience, and contain products of various sizes and shapes (Yam et al., 2005). Active Packaging is an innovative concept that can be defined as a mode of packaging in which the package, the product, and the environment interact to prolong shelf life or enhance safety or sensory properties, while maintaining the quality of the product (Suppakul et al., 2003). This type of technology is becoming more and more important for the food industry, involved in a globalized market in which a product produced in a country can be consumed in other countries. In addition there is a growing concern about foodborne diseases, and many companies are interested in the development of biosensors included in the packages in order to detect the presence of pathogens (Yam et al., 2005). As mentioned in Section 2, here we studied a large network for the current world trade (year 2011) of active & intelligent packaging for food industry, interconnecting categories like Country (CU), Company (CO), Product (PR), Food Type (FT), and product use or Packaging Type (PT) (Fig. 4). After calculating $\theta_k(i)$ and $\theta_k(j)$ for all pairs of connected nodes ($L_{ij}=1$) and for a large number of unconnected pairs in the network, we carried out a LDA. The best model found was

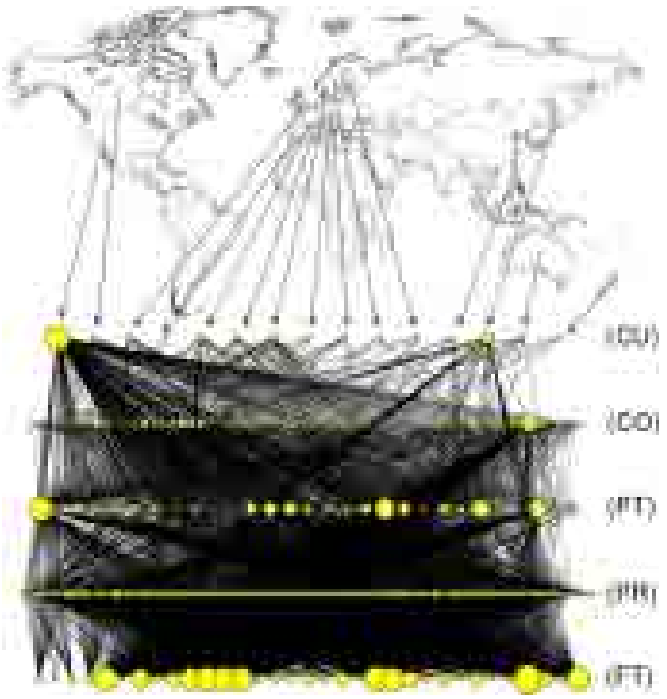


Fig. 4. Observed World Trade Network of Active & Intelligent Packaging for Food Industry. CU: country; CO: company; PT: packaging type; PR: product; FT: food type.

$$S(L_{ij}) = -2.00\theta_1(i) - 142.87\theta_1(j) + 116.65\theta_5(j) + 0.72$$
$$n = 31{,}911 \quad \chi^2 = 19{,}022 \quad p < 0.001 \qquad (13)$$

The model presents very good values of Accuracy, Sensitivity, and Specificity (see Table 1). The parameter $\theta_1(i)$ quantifies the information referred to the trading relationships of the ith node with its direct neighbors ($k=1$) in the world trade network. The parameter $\theta_1(j)$ quantifies the same information for the jth node and its direct neighbors ($k=1$). The parameter θ_5(j) quantifies the information referred to middle-long range trading relationships ($k=5$) in the trade network between the jth node and its neighbors of any class. As in the previous equations, the value of $\chi^2=19{,}022$ corresponds to a p-level <0.001, which indicates a significant discrimination between successful and not-observed trading relationships. In order to re-evaluate this kind of network using this equation it is necessary to introduce the values of $\theta_k(i)$ and $\theta_k(j)$ for the ith and jth nodes according to the following hierarchical order in i to j direction: Country (CU) → Company (CO) → Product (PR) → Packaging Type (PT) → Food Type (FT). Therefore, if we want to predict the expected success of a given CO to introduce a determined PT in the world trading network the ith node should represent the CO and the jth node the PT. In this equation, $S(L_{ij})$ is a real-valued output variable that scores the expected success if we want to establish a connection between the ith node (CU, CO, PR, etc.) and the jth node (PR, PT, FT) in the World Trade network of A&I Packaging for Food Industry.

3.7. Model 7: US FDA drug-target network

Study of drug–target interaction networks is an important topic in drug development (Yildirim et al., 2007; Lee et al., 2009; Mestres et al., 2009). However, determination of these interactions by means of experimental methods is both costly and time-consuming. Therefore it is interesting to develop mathematical

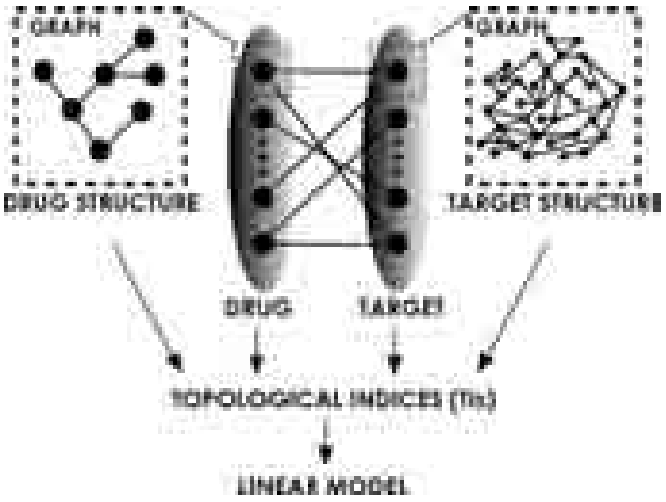


Fig. 5. Development of a linear model that takes into account the drug and target structure and the information about the drug/target nodes.

models in order to describe and evaluate in a simple way this type of interactions (He et al., 2010; Vina et al., 2009). Here, we have developed a model that takes into account the drug and target structure and the information about the drug/target nodes in the studied network (see Fig. 5). In this network $L_{ij}=1$ if the ith protein (T_i) is identified as target of the jth drug (D_j) in the DrugBank database and $L_{ij}=0$ otherwise. The best model found was

$$S(L_{ij}) = -1.10\cdot\theta_5(\text{D.node}) + 0.11\theta_0(\text{D.Total}) - 2.19\theta_3(\text{T.node}) - 0.47\theta_4(\text{T.middle}) - 1.43$$

$$n = 2{,}234 \quad \chi^2 = 2{,}123 \quad p < 0.001$$

where, θ_k is the entropy after k steps. The θ_k values used to seek the equation have been calculated for drugs (D.node) and targets (T.node) in the drug–target network. They were also calculated for the drug structure, taking into account all the atoms/nodes of the molecule/graph (D.Total) and for the target structure, taking into account the information codified in the amino acids/nodes from the middle of the protein/graph (T. Middle). To obtain this model, data were standardized previously to the LDA. The $\chi^2 = 2123$ statistics corresponds to a p-level <0.001, which indicates a significant discrimination between drug–target interaction and no interaction. The values of Accuracy, Sensitivity, and Specificity can be seen in Table 1. The data obtained from the model were used to re-construct the observed network and to compare both networks (giant components), by means of network descriptors (see Table 4).

4. Conclusions

In this work we confirm that it is possible to combine Markov Chains and Shannon Entropy in order to calculate higher order entropy parameters. We also show that these parameters can be used to quantify information about local and global node-node connections in different types of complex networks. For it, we have used MI-NODES, a new tool for the study of complex networks which is an upgrade of the software MARCH-INSIDE, classically used to study drugs and proteins. The parameters obtained can be used as inputs of LDA models to computationally assess the quality of connectivity patterns in known and new complex networks. These QSPR-like models are useful to re-construct and/or collate in a simple and cheap way a network, as alternative to high cost experimental re-evaluation of all links. This is a topic of the major importance because of the increasing use of existing complex networks in many areas of research. Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors (Chou and Shen, 2009), we shall make efforts in our future work to provide a web-server for the method presented in this paper.

Acknowledgments

Duardo-Sánchez, A. acknowledges partial financial support from Department of Especial Public Law, Financial and Tributary Law Area, Faculty of Law, University of Santiago de Compostela (Research Project (2006/PX 207). Xunta de Galicia and ESF). We also thank the Ibero-American Network of the Nano-Bio-Info-Cogno Convergent Technologies (Ibero-NBIC) network (209RT0366) funded by CYTED (Ciencia y Tecnología para el Desarrollo). González-Díaz H. and Munteanu C. R. acknowledge the Isidro Parga Pondal Programme, Xunta de Galicia, Spain.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2011.10.016.

References

Althaus, I.W., et al., 1993a. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 32, 6548–6554.
Althaus, I.W., et al., 1993b. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J. Biol. Chem. 268, 14875–14880.
Anderson, R.M., May, R.M., 1979. Population biology of infectious diseases: part I. Nature 280, 361–367.
Andraos, J., 2008. Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. Can. J. Chem. 86 (4), 342–357.
Arai, H.P., Mudry, D.R., 1983. Protozoan and metazoan parasites of fishes from the headwaters of the Parsnip and McGregor Rivers, British Columbia: a study of possible parasite transfaunations. Can. J. Fish. Aquat. Sci. 40, 1676–1684.
Arthur, J.R., Margolis, L., Arai, H.P., 1976. Parasites of fishes of Aishihik and Stevens Lakes, Yukon Territory, and potential consequences of their interlake transfer through a proposed water diversion for hydroelectrical purposes. J. Fish. Res. Board Can. 33, 2489–2499.
Baldazzi, V., et al., 2010. The carbon assimilation network in *Escherichia coli* is densely connected and largely sign-determined by directions of metabolic fluxes. PLoS Comput Biol 6 (6), e1000812.
Bangham, R.V., 1955. Studies on fish parasites of Lake Huron and Manitoulin Island. Am. Midl. Nat. 53, 184–194.
Batagelj, V., Mrvar, A., 1998. Pajek: a program for large network analysis. Connections 21 (2), 47–57.
Bird, D.M., Opperman, C.H., 1998. *Caenorhabditis elegans*: a genetic guide to parasitic nematode biology. J Nematol. 30 (3), 299–308.
Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U., 2006. Complex networks: structure and dynamics. Phys. Rep. 424, 175–308.
Bornholdt, S., Schuster, H.G., 2003. Handbook of Graphs and Complex Networks: From the Genome to the InternetWILEY-VCH GmbH & CO. KGaA, Wheinheim.
Burgard, A.P., Maranas, C.D., 2001. Probing the performance limits of the *Escherichia coli* metabolic network subject to gene additions or deletions. Biotechnol. Bioeng. 74 (5), 364–375.
Burglin, T.R., Lobos, E., Blaxter, M.L., 1998. *Caenorhabditis elegans* as a model for parasitic nematodes. Int. J. Parasitol. 28 (3), 395–411.
Chen, C., Chen, L., Zou, X., Cai, P., 2009. Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein Pept. Lett. 16, 27–31.
Chinniah, V.C., Threlfall, W., 1978. Metazoan parasites of fish from the Smallwood Reservoir, Labrador, Canada. J. Fish Biol. 13, 203–213.
Chou, K.C., 1989. Graphic rules in steady and non-steady state enzyme kinetics. J. Biol. Chem. 264 (20), 12074–12079.
Chou, K.C., 1990. Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady-state systems. Biophys. Chem. 35 (1), 1–24.
Chou, K.C., 2010. Graphic rule for drug metabolism systems. Curr Drug Metab. 11 (4), 369–378.
Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273 (1), 236–247.

Chou, K.C., Forsen, S., 1980. Graphical rules for enzyme-catalyzed rate laws. Biochem. J. 187, 829–835.
Chou, K.C., Shen, H.B., 2008. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nat. Protocols 3 (2), 153–162.
Chou, K.C., Shen, H.B., 2009. Recent advances in developing web-servers for predicting protein attributes. Nat. Sci. 2, 63–92.
Chou, K.C., Shen, H.B., 2010. Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. PloS ONE 5 (6), e11335.
Chou, K.C., Zhang, C.T., 1995. Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30 (4), 275–349.
Chou, K.C., Kezdy, F.J., Reusser, F., 1994. Review: steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. Anal. Biochem. 221, 217–230.
Chou, K.C., Zhang, C.T., Maggiora, G.M., 1997. Disposition of amphiphilic helices in heteropolar environments. Proteins 28 (1), 99–108.
Chou, K.C., Wu, Z.C., Xiao, X., 2011. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLoS ONE 6 (3), e18258.
Consortium TCeS, 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282 (5396), 2012–2018.Costa, R.S., Machado, D., Rocha, I., Ferreira, E.C., 2010. Hybrid dynamic modeling of *Escherichia coli* central metabolic network combining Michaelis-Menten and approximate kinetic equations. Biosystems 100 (2), 150–157.
De Nooy, W., Mrvar, A., Batagelj, V., 2005. Exploratory Social Network Analysis with PajekCambridge University Press, Cambridge (p 334).
Dechtiar, A.O., 1972. Parasites of fish from Lake of the Woods, Ontario. J. Fish. Res. Board Can. 29, 275–283.
Dehmer, M., Emmert-Streib, F. (Eds.), 2009. Analysis of Complex Networks: From Biology to Linguistics, Wiley-Blackwell, Wheinheim.
Dehmer, M., Varmuza, K., Borgert, S., Emmert-Streib, F., 2009. On entropy-based molecular descriptors: statistical analysis of real and synthetic chemical structures. Journal Chem. Inf. Modeling 49 (7), 1655–1663.
Desdevises, Y., Morand, S., Legendre, P., 2002. Evolution and determinants of host specificity in the genus Lamellodiscus (Monogenea). Biol. J. Linn. Soc. 77, 431–443.
Detwiler, J., Janovy Jr., J., 2008. The role of phylogeny and ecology in experimental host specificity: insights from a eugregarine-host system. J. Parasitol. 94 (1), 7–12.
Duardo-Sánchez, A., 2010. Study of criminal law networks with Markov-probability centralities. in: González-Díaz, H. (Ed.), Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks, Bentham, Kerala, India, pp. 205–212.
Duardo-Sánchez, A., 2011. Criminal law networks, markov chains, Shannon entropy and artificial neural networks. in: González-Díaz, H. (Ed.), Complex Network Entropy: From Molecules to Biology, Parasitology, Technology, Social, Legal, and Neurosciences, Bentham, Kerala, India, pp. 107–114.
Edwards, J.S., Palsson, B.O., 2000. Robustness analysis of the *Escherichia coli* metabolic network. Biotechnol. Prog. 16 (6), 927–939.
Fowler, J.H., Jeon, S., 2008. The authority of Supreme Court precedent. Soc. Networks 30, 16–30.
Fowler, Z.L., Gikandi, W.W., Koffas, M.A., 2009. Increased malonyl coenzyme A biosynthesis by tuning the *Escherichia coli* metabolic network and its application to flavanone production. Appl. Environ. Microbiol. 75 (18), 5831–5839.
Gerlee, P., Lizana, L., Sneppen, K., 2009. Pathway identification by network pruning in the metabolic network of *Escherichia coli*. Bioinformatics 25 (24), 3282–3288.
Ghim, C.M., Goh, K.I., Kahng, B., 2005. Lethality and synthetic lethality in the genome-wide metabolic network of *Escherichia coli*. J. Theor. Biol. 237 (4), 401–411.
Goffeau, A., 1997. The yeast genome directory. Nature 387 (6632, Suppl.), 5.
González-Díaz, H., Munteanu, C.R. (Eds.), 2010. Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks, Transworld Research Network Kerala (India).González-Díaz, H., et al., 2010. Network prediction of fasciolosis spreading in Galicia (NW Spain). in: González-Díaz, H., Munteanu, C.R. (Eds.), Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks, Transworld Research Network, Kerala (India), pp. 191–204.
Gonzalez-Diaz, H., et al., 2010. Review of MARCH-INSIDE and complex networks prediction of drugs: ADMET, anti-parasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers. Curr. Drug Metab. 11, 379–406.
Gu, Q., Ding, Y.S., Zhang, T.L., 2010. Prediction of G-protein-coupled receptor classes in low homology using Chou's pseudo amino acid composition with approximate entropy and hydrophobicity patterns. Protein Pept. Lett. 17 (5), 559–567.
Hatcher, J.M., Dick, J.T.A., Dunn, A.M., 2006. How parasites affect interactions between competitors and predators. Ecol. Lett. 9 (11), 1253–1271.
He, Z., et al., 2010. Predicting drug–target interaction networks based on functional groups and biological features. PLoS ONE 5 (3), e9603.
Hill, T., Lewicki, P., 2006. STATISTICS Methods and Applications: A Comprehensive Reference for Science, Industry and Data Mining. StatSoft, Tulsa p 813.
Holden-Dye, L., Walker, R.J., 2007. Anthelmintic drugs. WormBook, 1–13.
Huang, T., et al., 2010. Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. PLoS ONE 5 (6), e10972.
Huang, T., Chen, L., Cai, Y.D., Chou, K.C., 2011. Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property. PLoS ONE 6 (9), e25297.
Imielinski, M., Belta, C., Rubin, H., Halasz, A., 2006. Systematic analysis of conservation relations in Escherichia coli genome-scale metabolic network reveals novel growth media. Biophys. J. 90 (8), 2659–2672.
Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabasi, A.L., 2000. The large-scale organization of metabolic networks. Nature 407 (6804), 651–654.
Junker, B.H., Koschuetzki, D., Schreiber, F., 2006. Exploration of biological network centralities with CentiBiN. BMC Bioinf. 7 (1), 219.
Knox, C., et al., 2011. DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. Nucleic Acids Res. 39, 1035–1041.
Konig, R., et al., 2006. Discovering functional gene expression patterns in the metabolic network of *Escherichia coli* with wavelets transforms. BMC Bioinf. 7, 119.
Kotter, R., 2004. Online retrieval, processing, and visualization of primate connectivity data from the CoCoMac database. Neuroinformatics 2 (2), 127–144.
Lee, D.S., et al., 2008. The implications of human metabolic network topology for disease comorbidity. Proc. Natl. Acad. Sci. USA 105 (29), 9880–9885.
Lee, S., Park, K., Kim, D., 2009. Building a drug–target network and its applications. Expert Opin. Drug Discovery 4 (11), 1–13.
Leong, T.S., Holmes, J.C., 1981. Communities of metazoan parasites in open water fishes of Cold Lake, Alberta. J. Fish Biol. 18, 693–713.
Light, S., Kraulis, P., 2004. Network analysis of metabolic enzyme evolution in *Escherichia coli*. BMC Bioinf. 5, 15.
Lin, H., Bennett, G.N., San, K.Y., 2005. Chemostat culture characterization of *Escherichia coli* mutant strains metabolically engineered for aerobic succinate production: a study of the modified metabolic network based on metabolite profile, enzyme activity, and gene expression profile. Metab. Eng. 7 (5-6), 337–352.
Lin, W.Z., Fang, J.A., Xiao, X., Chou, K.C., 2011. iDNA-Prot: identification of DNA binding proteins using random forest with grey model. PLoS ONE 6 (9), e24756.
Mas-Coma, S., 2005. Epidemiology of fascioliasis in human endemic areas. J. Helminthol. 79 (3), 207–216.
Mestres, J., Gregori-Puigjane, E., Valverde, S., Sole, R.V., 2009. The topology of drug–target interaction networks: implicit dependence on drug properties and target families. Mol. Biosyst. 5 (9), 1051–1057.
Mezo, M., Gonzalez-Warleta, M., Castro-Hermida, J.A., Ubeira, F.M., 2008. Evaluation of the flukicide treatment policy for dairy cattle in Galicia (NW Spain). Vet. Parasitol. 157 (3–4), 235–243.
Modha, D.S., Singh, R., 2010. Network architecture of the long-distance pathways in the macaque brain. Proc. Natl. Acad. Sci. USA 107 (30), 13485–13490.
Mohabatkar, H., 2010. Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein Pept. Lett. 17, 1207–1214.
Muller, B., Grossniklaus, U., 2010. Model organisms—a historical perspective. J. Proteomics 73 (11), 2054–2063.
Newman, M., 2003. The structure and function of complex networks. SIAM Rev. 56, 167–256.
Nunn, C.L., Altizer, S.M., 2005. The global mammal parasite database: an online resource for infectious disease records in wild primates. Evol. Anthropol. 14, 1–2.
Pereira de Abreu, D.A., Cruz, J.M., Paseiro-Losada, P. Active and intelligent packaging for the food industry. Food Rev. Int. 27, doi:10.1080/87559129.2011.595022. In press.
Poulin, R., 2010. Network analysis shining light on parasite ecology and diversity. Trends Parasitol. 26 (10), 492–498.
Poulin, R., Krasnov, B.R., Mouillot, D., 2011. Host specificity in phylogenetic and geographic space. Trends Parasitol. 27 (8), 355–361.
Price, P.W., et al., 1986. Parasite mediation in ecological interactions. Annu. Rev. Ecol. Syst. 17, 485–505.
Puzyn, T., Leszczynski, J., Cronin, M.T.D. (Eds.), 2010. Recent Advances in QSAR Studies: Methods and Applications, Springer.Rosa da Silva, M., Sun, J., Ma, H.W., He, F., Zeng, A.P., 2008. Metabolic networks. in: Junker, B.H., Schreiber, F. (Eds.), Analysis of Biological Networks, Wiley & Sons, New Jersey, pp. 233–253.
Schmid, J.W., Mauch, K., Reuss, M., Gilles, E.D., Kremling, A., 2004. Metabolic design based on a coupled gene expression-metabolic network model of tryptophan production in *Escherichia coli*. Metab. Eng. 6 (4), 364–377.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423.
Shi, H., Nikawa, J., Shimizu, K., 1999. Effect of modifying metabolic network on poly-3-hydroxybutyrate biosynthesis in recombinant *Escherichia coli*. J Biosci. Bioeng. 87 (5), 666–677.
Stephan, K.E., et al., 2001. Advanced database methodology for the collation of connectivity data on the Macaque brain (CoCoMac). Philos. Trans. R. Soc. Lond. B. Biol. Sci 356 (1412), 1159–1186.
Suppakul, P., Miltz, J., Sonneveld, K., Bigger, S.W., 2003. Active packaging technologies with an emphasis on antimicrobial packaging and its applications. J. Food Sci. 68 (2), 408–420.
Vina, D., Uriarte, E., Orallo, F., Gonzalez-Diaz, H., 2009. Alignment-free prediction of a drug–target complex network based on parameters of drug connectivity and protein sequence of receptors. Mol. Pharmacol. 6 (3), 825–835.StatSoft.Inc., 2002. STATISTICA (data analysis software system), Version 6.0, 〈www.statsoft.com〉. Statsoft, Inc.), 6.0.
Wang, P., Xiao, X., Chou, K.C., 2011. NR-2L: a two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features. PLoS ONE 6 (8), e23505.
Wasserman, S., Faust, K., 1999. Social Network Analysis: Methods and ApplicationsCambridge University Press, Cambridge.

Wishart, D.S., 2010. DrugBank: a general resource for pharmaceutical and pharmacological research. Mol. Cell. Pharmacol. 2 (1), 25–38.
Wishart, D.S., et al., 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 34 (Database issue):D668-672.
Wishart, D.S., et al., 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 36 (Database issue):D901-906.
Wu, Z.C., Xiao, X., Chou, K.C., 2010. 2D-MH: a web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. J. Theor. Biol. 267 (1), 29–34.
Xiao, X., Wu, Z.C., Chou, K.C., 2011. iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. J. Theor. Biol. 284 (1), 42–51.
Xie, G., Mo, Z., 2011. Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications. J. Theor. Biol. 269 (1), 123–130.
Yam, K.L., Takhistov, P.T., Miltz, J., 2005. intelligent packaging: concepts and applications. J. Food Sci. 70 (1), 1–10.
Yildirim, M.A., Goh, K.I., Cusick, M.E., Barabasi, A.L., Vidal, M., 2007. Drug–target network. Nat. Biotechnol. 25 (10), 1119–1126.
Zakeri, P., Moshiri, B., Sadeghi, M., 2011. Prediction of protein submitochondria locations based on data fusion of various features of sequences. J. Theor. Biol. 269 (1), 208–216.
Zeng, Y.H., et al., 2009. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J. Theor. Biol. 259 (2), 366–372.
Zhang, J., Gao, D.Y., Yearwood, J., 2011. A novel canonical dual computational approach for prion AGAAAAGA amyloid fibril molecular modeling. J. Theor. Biol. 284 (1), 149–157.
Zhou, G.P., 2011a. The structural determinations of the leucine zipper coiled-coil domains of the cGMP-dependent protein kinase I alpha and its interaction with the myosin binding subunit of the myosin light chains phosphase. Protein Pept. Lett..
Zhou, G.P., 2011b. The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein–protein interaction mechanism. J. Theor. Biol. 284 (1), 142–148.
Zhou, G.P., Deng, M.H., 1984. An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. Biochem. J. 222, 169–176.

From QSAR models of Drugs to Complex Networks: State-of-Art Review and Introduction of New Markov-Spectral Moments Indices

Pablo Riera-Fernández[1], Raquel Martín-Romalde[1], Francisco J Prado-Prado[2], Manuel Escobar[2], Cristian R. Munteanu[3], Riccardo Concu[1], Aliuska Duardo-Sanchez[4] and Humberto González-Díaz[1,*]

[1]*Department of Microbiology & Parasitology, University of Santiago de Compostela (USC), Santiago de Compostela, 15782, Spain,* [2]*Department of Organic Chemistry, USC, Santiago de Compostela, 15782, Spain,* [3]*Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruña, Campus de Elviña, 15071, A Coruña, Spain,* [4]*Department of Especial Public Law, Financial and Tributary Law Area, Faculty of Law, USC, Santiago de Compostela, 15782, Spain*

Abstract: Quantitative Structure-Activity/Property Relationships (QSAR/QSPR) models have been largely used for different kind of problems in Medicinal Chemistry and other Biosciences as well. Nevertheless, the applications of QSAR models have been restricted to the study of small molecules in the past. In this context, many authors use molecular graphs, atoms (nodes) connected by chemical bonds (links) to represent and numerically characterize the molecular structure. On the other hand, Complex Networks are useful in solving problems in drug research and industry, developing mathematical representations of different systems. These systems move in a wide range from relatively simple graph representations of drug molecular structures (molecular graphs used in classic QSAR) to large systems. We can cite for instance, drug-target interaction networks, protein structure networks, protein interaction networks (PINs), or drug treatment in large geographical disease spreading networks. In any case, all complex networks have essentially the same components: nodes (atoms, drugs, proteins, microorganisms and/or parasites, geographical areas, drug policy legislations, *etc.*) and links (chemical bonds, drug-target interactions, drug-parasite treatment, drug use, etc.). Consequently, we can use the same type of numeric parameters called Topological Indices (TIs) to describe the connectivity patterns in all these kinds of Complex Networks irrespective the nature of the object they represent and use these TIs to develop QSAR/QSPR models beyond the classic frontiers of drugs small-sized molecules. The goal of this work, in first instance, is to offer a common background to all the manuscripts presented in this special issue. In so doing, we make a review of the most used software and databases, common types of QSAR/QSPR models, and complex networks involving drugs or their targets. In addition, we review both classic TIs that have been used to describe the molecular structure of drugs and/or larger complex networks. In second instance, we use for the first time a Markov chain model to generalize Spectral moments to higher order analogues coined here as the Stochastic Spectral Moments TIs of order k (π_k). Lastly, we report for the first time different QSAR/QSPR models for different classes of networks found in drug research, nature, technology, and social-legal sciences using π_k values. This work updates our previous reviews González-Díaz *et al.* Curr Top Med Chem. 2007; 7(10): 1015-29 and González-Díaz *et al.* Curr Top Med Chem. 2008; 8(18):1676-90. It has been prepared in response to the kind invitation of the editor Prof. AB Reitz in commemoration of the 10th anniversary of this journal in 2010.

Keywords: Centralities, Complex networks, Markov chains, QSAR, QSPR, Spectral moments, Topological indices.

INTRODUCTION

Quantitative Structure-Activity/Property Relationships (QSAR/QSPR) models have been largely used for different kind of problems in Medicinal Chemistry and other Biosciences as well. Nevertheless, the applications of QSAR models have been restricted to the study of small molecules in the past. In this context, many authors use molecular graphs, atoms (nodes) connected by chemical bonds (links) to represent and numerically characterize the molecular structure. However, more recently have appeared many QSAR/QSPR models with applications to more general situations. For instance, new QSAR/QSPR models may be applied to predict the function of a protein with a given sequence or 3D structure, the function of an RNA secondary structure, the interactions of specific drugs with multiple targets selected out of a large number of possible proteins present in the proteome of one organism or multiple infectious/parasitic organisms [1, 2]. In this sense, the editor González-Díaz has recently organized series of special issues with both the old and new applications of QSAR, which have been published on Current Topics in Medicinal Chemistry [2-11], Current Proteomics [12-19], Current Drug Metabolism [20-28], Current Pharmaceutical Design [29-38], and Current Bioinformatics [39-48].

In all these reviews we can note that, Graph and Complex Network theory is expanding their applications to different

*Address correspondence to this author at Department of Microbiology & Parasitology, Faculty of Pharmacy, USC, 15782, Santiago de Compostela, Spain, Tel: +34 881 814 49 85; Fax: + 34 981594912,
Email: gonzalezdiazh@yahoo.es

1873-5294/12 $58.00+.00

levels of matter organization such as the genome networks, protein-protein networks, sexual disease transmission networks, linguistic networks, low and social networks [49-54], power electric power network or internet [55]. A network is a set of items, usually called *nodes*, with connections between them, so called *edges* [54]. The nodes can be atoms, molecules, proteins, nucleic acids, drugs, cells, organisms, parasites, people, words, laws, computers or any other part of a real system. The edges are relationships between the nodes such as chemical bonds, physical interactions, metabolic pathways, pharmacological action, law recurrence or social behavior (see also reviews cited above).

In this work, we offer a common background to all the manuscripts presented in this special issue. In so doing, we make a review of the most used software and databases, common types of QSAR/QSPR models, and complex networks involving drugs or their targets. In addition, we review both classic TIs that have been used to describe the molecular structure of drugs and/or larger complex networks. In second instance, we use for the first time a Markov chain model to generalize classic spectral moments to higher order analogues coined here as the Stochastic Spectral Moments of order k (π_k). Lastly, we report for the first time different QSAR/QSPR models for different classes of networks found in drug research, nature, technology, and social-legal sciences using π_k values. The list of topics for the present work is:

1. DATABASES FOR QSAR

In this section, we are going to cite some of the most common databases used in QSPR/QSAR studies related with drug discovery.

1.1. Protein Data Bank (PDB)

The PDB was established in 1971 at Brookhaven National Laboratory as an archive for biological macromolecular crystal structures. In 1998, the Research Collaboratory for Structural Bioinformatics (RCSB) (http://www.rcsb.org/pdb/home/home.do) became responsible for the management of the PDB and in 2003 the worldwide PDB (wwPDB) (http://www.wwpdb.org/) was established to formally recognize the international nature of the PDB archive and to ensure that the data files remain uniform in content and format. The founding members are the RCSB PDB (USA), the Macromolecular Structure Database at the European Bioinformatics Institute (MSD-EBI) (http://www.ebi.ac.uk/pdbe/) and the Protein Data Bank Japan (PDBj) (http://www.pdbj.org/) at Osaka University. These wwPDB sites share responsibilities in data deposition, processing and distribution of the PDB archive, and agree to support a single, standardized archive of structural data. The BioMagResBank (BMRB) (http://www.bmrb.wisc.edu/) at the University of Wisconsin-Madison (USA) became a member in 2006 [56-58]. Currently, The PDB archive contains information about more than 73,000 experimentally-determined structures of proteins, nucleic acids, and complex assemblies. The PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules can be visualized and downloaded.

1.2. Drug Bank

The Drug Bank database is a bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. The first version of DrugBank (http://www.drugbank.ca/) was released in 2006 [59], the second in 2008 [60, 61] and the third in 2010 [62]. Currently it contains over 6,800 drug entries including >1,400 FDA-approved small molecule drugs, 134 FDA-approved biotech (protein/peptide) drugs, 83 nutraceuticals and >5,200 experimental drugs. Additionally, more than 4,400 non-redundant protein (i.e. drug target) sequences are linked to these FDA approved drug entries. Each DrugCard entry contains more than 150 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data. DrugBank is supported by David Wishart, Departments of Computing Science & Biological Sciences, University of Alberta.

1.3. ChEMBL

ChEMBL is a database of bioactive drug-like small molecules maintained by the European Bioinformatics Institute (EBI) (https://www.ebi.ac.uk/chembl/). The database, originally known as StARlite, was developed by a pharmaceutical company, Galapagos NV. It was acquired by the European Molecular Biology Laboratory (EMBL) in 2008 with an award from The Wellcome Trust, resulting in the creation of the ChEMBL chemogenomics group at EBI, led by John Overington [63, 64]. The database contains 2-D structures, calculated properties (e.g. logP, Molecular Weight, Lipinski Parameters, etc.) and abstracted bioactivities (e.g. binding constants, pharmacology and ADMET data). The authors attempt to normalize the bioactivities into a uniform set of end-points and units where possible, and also to tag the links between a molecular target and a published assay with a set of varying confidence levels. The data are abstracted and curated from the primary scientific literature, and cover a significant fraction of the Structure Activity Relationships (SAR) and discovery of modern drugs. Additional data on clinical progress of compounds is being integrated into ChEMBL at the current time. Version 9 contains 757,845 compound records (658,075 distinct compounds of which 657,736 have molfiles); 499,867 assays 3,030,317 activities and 8,091 targets. Data are abstracted from a total of 39,094 publications.

2. SOFTWARE FOR MOLECULAR PARAMETERS

Molecular descriptors play a fundamental role in QSPR/QSAR studies. They can be defined as the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment [65]. In this section we are going to present some programs used to calculate molecular descriptors.

2.1. Dragon

DRAGON(http://www.talete.mi.it/products/dragon_description.htm) software has been conceived to provide the user with a variety of molecular descriptors derived from differ-

Table 1. Dragon (v. 6.0) Families of Molecular Descriptors

ID Block	Block Description	N°
1	Constitutional descriptors	43
2	Ring descriptors	32
3	Topological indices	75
4	Walk and path counts	46
5	Connectivity indices	37
6	Information indices	48
7	2D matrix-based descriptors	550
8	2D autocorrelations	213
9	Burden eigenvalues	96
10	P_VSA-like descriptors	45
11	ETA indices	23
12	Edge adjacency indices	324
13	Geometrical descriptors	38
14	3D matrix-based descriptors	90
15	3D autocorrelations	80
16	RDF descriptors	210
17	3D-MoRSE descriptors	224
18	WHIM descriptors	114
19	GETAWAY descriptors	273
20	Randic molecular profiles	41
21	Functional group counts	154
22	Atom-centred fragments	115
23	Atom-type E-state indices	170
24	CATS 2D	150
25	2D Atom Pairs	1,596
26	3D Atom Pairs	36
27	Charge descriptors	15
28	Molecular properties	20
29	Drug-like indices	27

ent molecular representations. The first release of DRAGON was developed in 1994 by Milano Chemometrics and QSAR Research Group with the name WHIM/3D QSAR, being specific for the calculation of the WHIM descriptors. Successively, a lot of other descriptors have been implemented leading to a new software, which in 1997 provided about 600 descriptors and was released with the name DRAGON [66]. Currently, DRAGON (v. 6.0) allows the calculation of 4,855 molecular descriptors divided into 29 blocks (Table **1**) and it is managed by TALETE S.R.L., a commercial brand. E-DRAGON (v. 1.0) (http://www.vcclab.org/lab/edragon/) is the electronic remote version of DRAGON (v. 5.4). It is free and allows the calculation of more than 1,600 molecular descriptors that are divided into 20 logical blocks [67]. E-

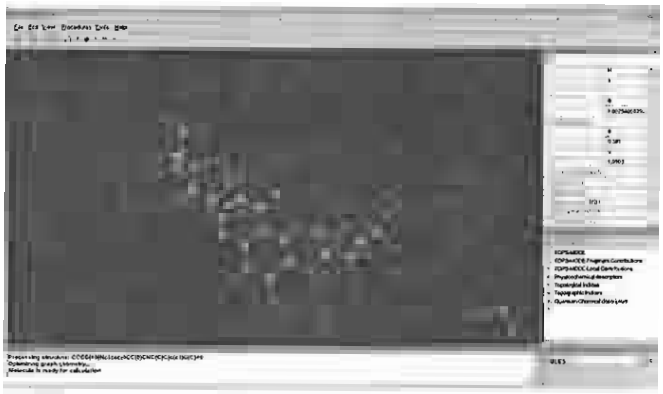


Fig. (1). MODesLab graphical interface.

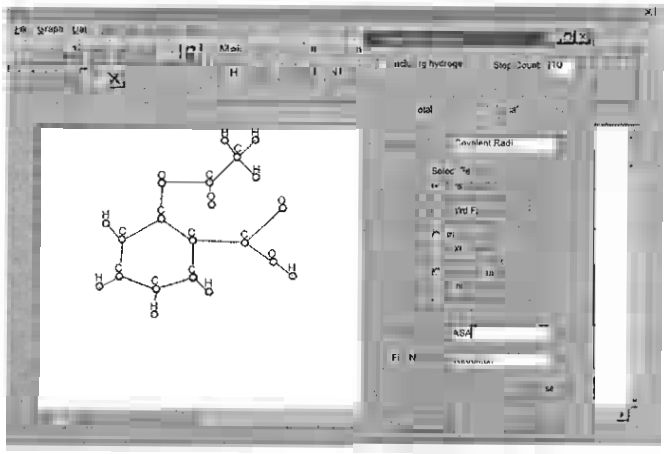


Fig. (2). TOMOCOMD graphical interface.

Dragon was developed as a result of collaboration between Dr. Tetko, Prof. Todeschini's and Prof. Gasteiger's teams. Some examples in the literature of the use of dragon software are [68-70].

2.2. MoDes Lab

MoDes Lab (http://www.modeslab.com/), developed by E. Estrada and Y. Gutierrez, was released for the first time in 2002. Currently we can find the version 1.5, released in 2004 (Fig. **1**). It provides all the necessary tools to perform QSAR studies, from the input of large number of molecules to the calculations of molecular descriptors (e.g. Kier & Hall, Kappa and Balaban indices, Abraham descriptors and TopsMode sub-structural descriptors), property prediction and substructural analysis. It also provides a very useful way to define the properties of atoms, bonds and fragments by an extension of SMILES language and use these properties in molecular descriptors calculations. Some examples of the use of this program in research are [71-73].

2.3. Tomocomd

In 2002 Y. Marrero-Ponce & V. Romero released the version 1.0 of TOMOCOMD (TOpological MOlecular COMputer Design) (Fig. **2**). It consists of 4 subprograms and every one of them allows both drawing the structures (drawing mode) and calculating molecular 2D/3D descriptors (calculation mode). The modules are named CARDD (Computed-Aided 'Rational' Drug Design), CAMPS (Computed-

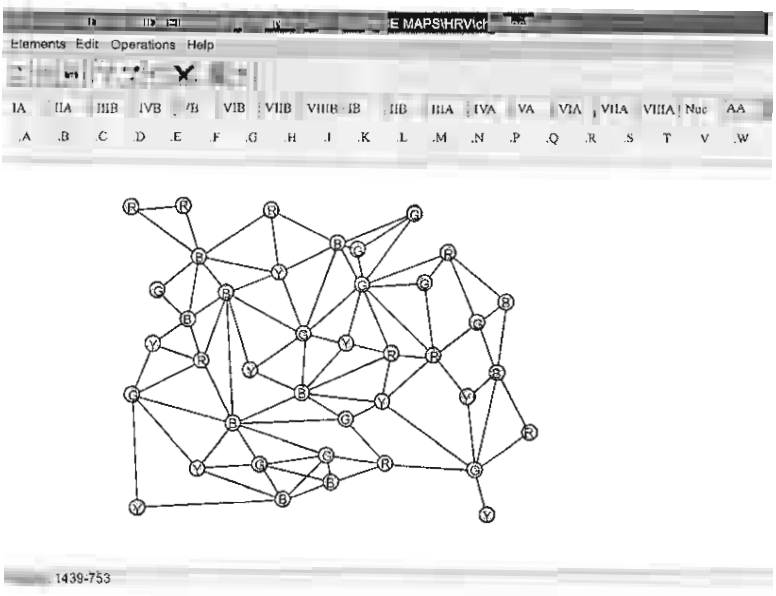


Fig. (3). MARCH-INSIDE graphical interface.

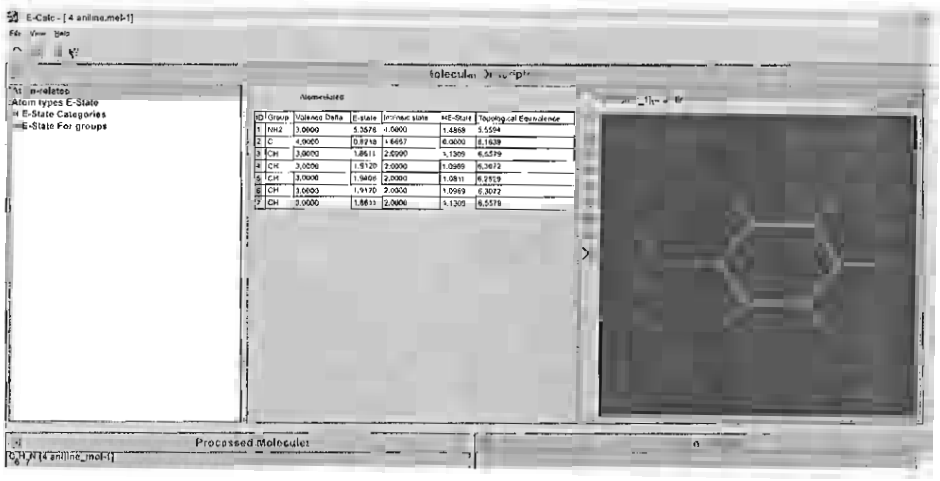


Fig. (4). E-Calc graphical interface.

Aided Modeling in Protein Science), CANAR (Computed-Aided Nucleic Acid Research), and CABPD (Computed-Aided Bio-Polymers Docking) [74]. The software calculates different types of topological indices including quadratic $q_k(\mathbf{w})$, linear $f_k(\mathbf{w})$ and bi-linear $b_k(\mathbf{w},\mathbf{v})$ forms. A review of the aplications of TOMOCOMD in QSPR/QSAR studies of antiparasitic drugs can be found in [35].

2.4. March-Inside

MARCH-INSIDE (MARkovian CHemicals IN SIlico DEsign) is a simple but efficient computational approach to the study of QSAR in medicinal chemistry developed by Gonzalez-Díaz *et al.* (Fig. **3**). It uses the theory of Markov Chains to generate parameters that numerically describe the chemical structure of drugs and drug targets. This approach generates two principal types of parameters: Stochastic Topological Indices and stochastic 3D-Topographic Indices. In the following reviews we can find examples of the use of MARCH-INSIDE in studies about prediction of antimicrobial/antiparasite agents as well as their molecular targets [10, 35, 75].

2.5. E-Calc

E-Calc (v.1.1/1999) is an utility included in the book *Molecular Structure Description: The Electrotopological State* [76] that calculates E-State values for molecules, including the electrotopological state (E-State) and hydrogen E-State (HE-State) values for individual atoms as well as the atom-type E-State indices (Fig. **4**). These calculations help to understand the development, use, and interpretation of E-State values as a representation of molecular structure. The computational parts of this program have been taken from Molconn-Z and from SciQSAR 2D.

2.6. Codessa Pro

CODESSA PRO (Comprehensive Descriptors for Structural and Statistical Analysis) (http://www.codessa-pro.com/) is a program desingned by Alan R. Katritzky, Mati Karelson and Ruslan Petrukhin and developed from 2001 to 2005. According to the user's manual (http://www.codessa-pro.com/manuals/manual.htm) it is designed for developing quantitative structure-activity/property relationships

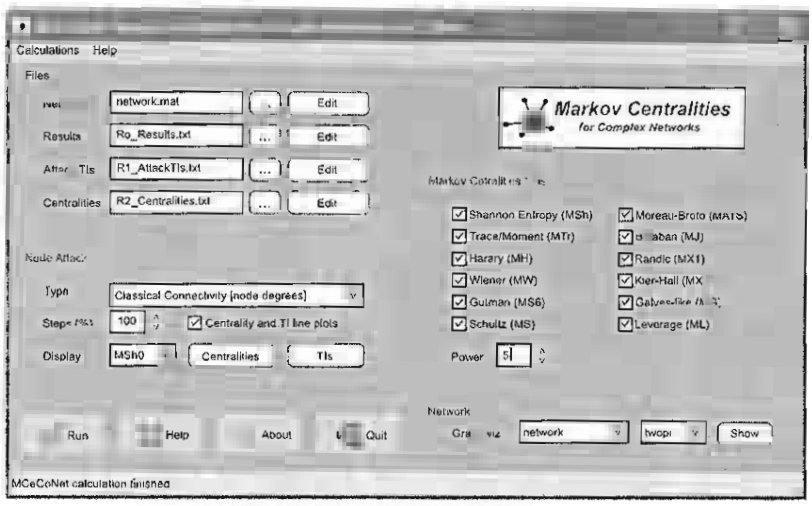


Fig. (5). MCeCoNet graphical interface.

(QSAR/QSPR) by integrating all necessary mathematical and computational tools to: (i) calculate a large variety of molecular descriptors on the basis of the 3D geometrical structure and/or quantum-chemical wave function of chemical compounds; (ii) develop (multi)linear and non-linear QSPR models on the chemical and physical properties or biological activity of chemical compounds; (iii) carry out cluster analysis of the experimental data and molecular descriptors; (iv) interpret the developed models; and (v) predict property values for any chemical compound with known molecular structure. CODESSA PRO includes 116 molecular descriptors divided into 8 groups (constitutional, topological, geometrical, electrostatic, CPSA, quantum-chemical, MO-related and thermodynamic). Some examples of the use of this program in research are [77-80].

3. SOFTWARE FOR NETWORK PARAMETERS

Many phenomena can be modeled as complex networks and, as we will see in the next sections, the network theory can be used in studies about drug discovery, metabolic pathways, diseases... Because of this, in this section we are going to present some programs dedicated to the analysis and graphical representation of networks.

31. Pajek

Pajek (Slovene word for spider) is a program for analysis of large networks developed by Vladimir Batagelj and Andrew Mrvar (with contributions of Matjaž Zaveršnik). The first version was released in 1996 and currently we can find the version 2.02, released in 2010. It is freely available, for noncommercial use, at its download page (http:// pajek.imfm.si/doku.php?id=pajek). According to the authors [81], the main goals of the design of Pajek are: to support abstraction by (recursive) factorization of a large network into several smaller networks that can be further treated using more sophisticated methods, to provide the user with some powerful visualisation tools and to implement a selection of efficient algorithms for analysis of large networks. A good source of information about pajek and its possibilities is the book *Exploratory social network analysis with Pajek* [82].

3.2. Cytoscape

Cytoscape (http://www.cytoscape.org/) is an open source software platform for visualizing complex networks and integrating these with any type of attribute data. A lot of plugins are available for various kinds of problem domains, including bioinformatics, social network analysis, and semantic web. An interesting feature of this program is that it can directly connect to the external public databases and imports network and annotation data. Cytoscape was initially made public in July, 2002 (v 0.8) and currently we can download the version 2.8.1. It is developed by the Institute for Systems Biology (Leroy Hood lab), the University of California San Diego (Trey Ideker lab), Memorial Sloan-Kettering Cancer Center (Chris Sander lab), the Institut Pasteur (Benno Schwikowski lab), Agilent Technologies (Annette Adler lab) and the University of California, San Francisco (Bruce Conklin lab). More information about Cytoscape can be found in [83-85].

3.3. CentiBiN

CentiBiN (Centralities in Biological Networks) (http://centibin.ipk-gatersleben.de/index.php) is a free tool for the computation and exploration of centralities in biological networks developed by Dirk Koschützki (Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)). It was released for the first time in 2004 and currently we can find the version 1.4.3, released in 2007. It supports a wide range of different centrality measures ranging from local measures (which only consider the direct neighbourhood of a vertex) to global measures. In total 17 centralities for undirected networks and 15 centralities for directed networks are available. CentiBin can draw an existing network or generate a random network by using one of the 5 algorithms available. Another interesting feature is that the distribution of centrality values and a histogram of centrality values can be displayed. [86]. Some examples of the use of CentiBin in research can be find in [87-89].

3.4. MCeCoNet

MceConet (Markov Centralities for Complex Networks) (Fig. **5**) is a program developed by H. González-Díaz and

C.R. Munteanu in 2009 (V. 1.0) that introduces a new class of centralities based in the Markov topological indices and node transition probabilities. An interesting feature is that it calculates these topological indices and node centralities during a network attack. The nodes are eliminated in the order of *classical connectivity* (node degrees), *average probability* for each node [avg(p_{ij})], *entropy* of each node [-p_j*log(p_j)], node *asymmetries* as difference between the averaged probability by line and column [p_{ij}-p_{ji}], *autocorrelation* as the product of the same averaged probabilities [p_{ij}*p_{ji}] or by keeping the *original matrix* (without sorting). MCeCoNet can calculate the following types of centralities and topological indices: Markov Shannon Entropy, Markov Traces, Markov Harary number, Markov Wiener index, Markov Gutman topological index, Markov Schultz topological index, Markov Moreau-Broto indices, Markov Balaban distance connectivity index, Markov Kier-Hall connectivity indices, Markov Randic connectivity index, Markov Galves indices and Markov Leverage indices. The networks are displayed by using several types of drawing *applications* (*dot*, *circo*, *twopi*, *neato* and *fdp* from *Graphviz*) and the plots corresponding to centralities/TIs for an attack are generated using *gnuplot*. An example of its use can be found in [47].

4. MULTI QSPR/QSAR MODELS FOR DRUG NETWORK ASSEMBLE

One limitation of almost all QSPR/QSAR models is that they predict the biological activity of drugs against only one biological system (organism, target...). Therefore, the development of multi task QSPR/QSAR models (mt-QSPR/mt-QSAR) to predict drugs activity/properties against different biological systems is an interesting field of study. These mt-QSPR/QSARs offer also a good opportunity to construct complex networks that can be used to explore large and complex drug-biological system databases. In this section we are going to review some of the mt-QSPR/QSAR models proposed in the literature and the networks derived from these studies.

4.1. mt-QSAR for Anti-Viral Drugs

Prado-Prado *et al.* [90] used the Markov Chain theory to calculate new multi-target spectral moments to fit a mt-QSAR model for drugs active against 40 viral species. The model is based on 500 drugs (including active and non-active compounds) tested as antiviral agents in the recent literature; not all drugs were predicted against all viruses, only those with experimental values. The database also contains 207 well-known compounds (which are not as recent as the previous ones). These compounds have been reported in the Merck Index with other activities that do not include antiviral action against any virus species and have been used as non-active compounds. Linear Discriminant Analysis (LDA) was used to classify all these drugs into two classes as active or non-active against the different viral species tested. The model correctly classified 5,129 out of 5,594 non-active compounds (sensitivity = 91.69 %) and 412 out of 422 active compounds (specificity = 97.63 %). Overall training predictability was 92.34 % (accuracy). Validation of the model was carried out by means of external predicting series, being classified 2,568 out of 2,779 (sensitivity = 92.41 %) non-active compounds and 224 out of 229 (specificity = 97.82 %) active compounds. Overall training predictability was 92.82 % (accuracy). The equation of the model is the following:

$$\begin{aligned} Actv = & -0.95 \cdot {}^{0}\mu_s(\text{H-Het}) + 1.50 \cdot {}^{2}\mu_s(\text{H-Het}) - 3.23 \cdot {}^{0}\mu_s(\text{C}_{uns}) \\ & - 4.02 \cdot {}^{0}\mu_s(\text{C}_{sat}) - 0.47 \cdot {}^{1}\mu_s(\text{T}) + 10.34 \cdot {}^{0}\mu_s(\text{T}) + 0.74 \cdot {}^{5}\mu_s(\text{X}) \\ & - 8.88 \end{aligned}$$

$$\lambda = 0,51; \quad \chi^2 = 4024.83; \quad p < 0.001$$

Where, λ is the Wilk's statistic; χ^2 chi square and p the error level. In the equation, ${}^{k}\mu_s$ is the spectral moment for a given specie after k steps. It has been calculated for the total (T) of atoms in the molecule or for specific collections of atoms. These collections are atoms with a common characteristic: H-Het: hydrogen bound to heteroatom, C_{uns}: unsaturated carbon atoms, C_{sat}: saturated Carbon atoms, X: halogen atoms.

Prado-Prado *et al.* [89] used the LDA to fit a mt-QSAR model that classified 600 drugs as active or non-active against the 41 different tested species of virus. The model correctly classified 143 out of 169 active antiviral compounds (specificity = 84.62 %) and 119 out of 139 non-active compounds (sensitivity = 85.61 %). Overall training accuracy was 85.1 % (262 out of 308 cases). On the other hand, validation of the model was carrying out by means of external predicting series, obtaining a cross validation accuracy of 90.7 % (466 out of 514 compounds). In order to illustrate the performance of the model in practice, it was developed a virtual screening that recognized as active 102 out of 110 (92,7 %) antiviral compounds not used in training or predicting series. The equation of the model is the following:

$$\begin{aligned} Actv = & 1.90 \cdot {}^{0}C_s(\text{C}_{sat}) - 1.64 \cdot {}^{0}C_s(\text{C}_{uns}) + 1.02 \cdot {}^{2}C_s(\text{C}_{uns}) \\ & + 1.10 \cdot {}^{5}C_s(\text{C}_s) + 0.73 \cdot {}^{1}C_s(\text{X}) + 1.08 \cdot {}^{1}C_s(\text{Het}) \\ & + 1.07 \cdot {}^{0}C_s(\text{H-Het}) - 0.75 \cdot {}^{4}C_s(\text{H-Het}) + 0.08 \end{aligned}$$

$$\lambda = 0.47; \quad Rc = 0.726; \quad p < 0.001$$

Where, λ is the Wilk's statistic; Rc is the canonical correlation and p the error level. In the equation ${}^{k}C_s$ is the molecular index for a given specie after k steps. It has been calculated for the total (T) of atoms in the molecule or for specific collections of atoms. These collections are atoms with a common characteristic: H-Het: hydrogen bound to heteroatom, C_{uns}: unsaturated carbon atoms, C_{sat}: saturated Carbon atoms, X: halogen atoms.

The authors also obtained and compared the topology of the complex networks and their respective giant components (GCs) based on different distance measures (Euclidean, Manhattan, Chebychey, Pearson, Percent disagreement and Power sum). It was observed that the obtained topological values varied notably from one network to other, being possible to classify them in two groups. The first group is composed of the Manhattan and Percent disagreement networks, with small-world features. The second group is composed of the other networks, with similar features to the Low-Density Random networks. The GC of the Manhattan network showed the more interesting features for drug–drug similarity search. In this work it is also given the procedure for the construction of back-projection maps for the contribution of each drug sub-structure to the antiviral activity against different species.

4.2. mt-QSAR for Anti-Bacterial Drugs

Prado-Prado *et al.* [91] developed a Markov model to describe the biological activity of more than 70 drugs from the literature tested against 96 species of bacteria. They applied LDA to classify drugs as active or inactive against the different tested bacterial species. The model correctly classified 199 out of 237 active compounds (83.9 %) and 168 out of 200 inactive compounds (84 %). Overall training predictability was 84 % (367 out of 437 cases). Validation of the model was carried out by means of external predicting series, being classified correctly 202 out of 243 (83.13 %) cases. In order to show how the model functions in practice a virtual screening was carried out, recognizing the model as active 480 out of 568 (84.5 %) antibacterial compounds not used in the training or predicting series. The equation of the model is the following:

$$\begin{aligned} Actv = & -1.12 \cdot {}^{1}C_s(\mathrm{T}) + 1.34 \cdot {}^{3}C_s(\mathrm{T}) + 1.84 \cdot {}^{0}C_s(\mathrm{C_{sat}}) \\ & - 0.90 \cdot {}^{0}C_s(\mathrm{C_{uns}}) + 0.88 \cdot {}^{5}C_s(\mathrm{X}) - 1.27 \cdot {}^{0}C_s(\mathrm{H\text{-}Het}) \\ & - 0.90 \cdot {}^{2}C_s(\mathrm{H\text{-}Het}) + 0.698 \\ & \lambda = 0.49; \quad \mathrm{Rc} = 0.715; \quad p < 0.001 \end{aligned}$$

Where, λ is the Wilk's statistic; Rc is the canonical correlation and p the error level. In the equation ${}^{k}C_s$ is the molecular index for a given specie after k steps. It has been calculated for the total (T) of atoms in the molecule or for specific collections of atoms. These collections are atoms with a common characteristic: H-Het: hydrogen bound to heteroatom, C_{uns}: unsaturated carbon atoms, C_{sat}: saturated Carbon atoms. X: halogen atoms.

Another model, proposed by Prado-Prado *et al.* [92] correctly classified 202 out of 241 active compounds (83.8 %) and 169 out of 200 non-active cases (84.5 %). Overall training predictability was 84.13 % (371 out of 441 cases). Validation of the model was carried out by means of external predicting series, being classified correctly 197 out of 221 (89.4 %) cases. In order to show how the model functions in practice a virtual screening was carried out, recognizing the model as active 86.7 %, 520 out of 600 cases not used in training or predicting series. The equation of the model is the following:

$$\begin{aligned} Actv = & -3.5 \cdot \pi_1(\mathrm{C_{sat}}) + 3 \cdot \pi_0(\mathrm{C_{sat}}) + 1.76 \cdot \pi_2(\mathrm{C_{uns}}) - 1.77 \cdot \pi_3(\mathrm{Het}) \\ & + 2.54 \cdot \pi_5(\mathrm{H\text{-}Het}) + 2.4 \cdot \pi_3(\mathrm{Het\text{-}Het}) - 5.42 \cdot \pi_2(\mathrm{H\text{-}Het}) \\ & + 0.74 \\ & \lambda = 0.49; \quad \mathrm{Rc} = 0.718; \quad p < 0.001 \end{aligned}$$

Where, λ is the Wilk's statistic; Rc the canonical index and p the error level. In the equation, π_k is the spectral moment for a given specie after k steps. It has been calculated for the total (T) of atoms in the molecule or for specific collections of atoms. These collections are atoms with a common characteristic: H-Het: hydrogen bound to heteroatom, C_{uns}: unsaturated carbon atoms, C_{sat}: saturated Carbon atoms, X: halogen atoms.

The outputs of this QSAR model were used as inputs to construct a network. The observed network has 1,242 nodes (drugs and bacteria), 772,736 edges (drug-bacteria pairs) with similar activity. The predicted network has 1,031 nodes and 641,377 edges. After edge-to-edge comparison, it has been demonstrated that the predicted network is significantly similar to the observed one and both have distribution closer to exponential than to normal.

4.3. mt-QSAR for Anti-Parasite Drugs

Prado-Prado *et al.* [88] proposed a mt-QSAR for more than 500 drugs tested in the literature against different parasites. The data were processed by LDA, classifying drugs as active or non-active against the different tested parasite species. The model correctly classified 212 out of 244 (87.0 %) cases in training series and 207 out of 243 compounds (85.4 %) in external validation series. In order to illustrate the performance of the QSAR for the selection of active drugs it was carried out an additional virtual screening of antiparasite compounds not used in training or predicting series. The model recognized 97 out of 114 (85.1 %) of them. The equation of the model is the following:

$$\begin{aligned} Actv = & 4.15 \times 10^{-14} \cdot {}^{1}C_s(\mathrm{T}) + 8.9 \times 10^{-14} \cdot {}^{0}C_s(\mathrm{C_{sat}}) \\ & - 1.5 \times 10^{-13} \cdot {}^{0}C_s(\mathrm{C_{uns}}) + 4.7 \times 10^{-7} \cdot {}^{5}C_s(\mathrm{C_{uns}}) \\ & + 2 \times 10^{-7} \cdot {}^{0}C_s(\mathrm{Het}) - 7.9 \times 10^{-7} \cdot {}^{4}C_s(\mathrm{H\text{-}Het}) - 0.72 \\ & \mathrm{Rc} = 0{,}75 \quad \lambda = 0.434; \quad \mathrm{F} = 51{,}44; \quad p < 0.001 \end{aligned}$$

Where, Rc is the canonical correlation coefficient, λ is the Wilk's statistic, F is the Fisher ratio and p the error level. In this equation ${}^{k}C_s$ is the molecular index for a given specie after k steps. It has been calculated for the total (T) of atoms in the molecule or for specific collections of atoms. These collections are atoms with a common characteristic: Het: heteroatoms, H-Het: hydrogen bound to heteroatom, C_{uns}: unsaturated carbon atoms, C_{sat}: saturated carbon atoms.

The authors also give the procedures to construct back-projection maps and to calculate sub-structures contribution to the biological activity. They used the outputs of the model to construct a multi-species complex network of antiparasite drugs. The network predicted has 380 nodes (compounds) and 634 edges (pairs of compounds with similar activity).

Prado-Prado *et al.* [93] developed a mt-QSAR model for more than 700 drugs tested in the literature against different parasites (predicting antiparasitic drugs). The data were processed by LDA and the model classified correctly 93.62 % (1,160 out of 1,239 cases) in training. Validation of the model was carried out by means of external predicting series, being classified correctly 573 out of 607 (94.4 %) cases. The equation of the model is the following:

$$\begin{aligned} Actv = & -3.86 \cdot {}^{A}\pi_1(\mathrm{s,C_{sat}}) - 3.71 \cdot {}^{A}\pi_1(\mathrm{s,C_{sp\&sp2}}) - 53.55 \cdot {}^{A}\pi_1(\mathrm{s,X}) \\ & + 50.92 \cdot {}^{A}\pi_3(\mathrm{s,X}) - 2.62 \cdot {}^{A}\pi_1(\mathrm{s,H\text{-}Het}) + 3.12 \cdot {}^{A}\pi_s(\mathrm{s,H-Het}) \\ & - 2.37 \\ & \mathrm{Rc} = 0.73; \ \lambda = 0.46; \quad p < 0.001 \end{aligned}$$

Where, Rc is the canonical correlation coefficient, λ is the Wilk's statistics and p is the error level. In this equation the absolute probabilities ${}^{A}\pi_k$ calculated refer to: ${}^{A}\pi_{0,1}$(s, $C_{sp\&sp2}$): all unsaturated carbon atoms (sp and sp2 atoms) and all atoms placed at distance d=5 from them. ${}^{A}\pi_1$(s, C_{sat}): all saturated carbon atoms. ${}^{A}\pi_1$(s, X): all halogen atoms. ${}^{A}\pi_0$(s, H-Het): all Hydrogen atoms bound to a Heteroatom (N, O, or S).

By using this model it is possible to construct drug–drug multispecies Complex Networks (msCN) and species–species multi drug resistance Complex Networks (mdrCN). The authors carried out a comparative study of the topology of six different drug–drug msCN based on six different distances such as Euclidean, Chebychev, Manhattan, Pearson, Percent disagreement and Power sum. Furthermore, they compared the selected drug–drug msCN and species–species mdrCN with random networks. Lastly, they reported the first substructural analysis of drug–drug msCN using Triadic Census Analysis algorithm.

Prado-Prado *et al.* [94] used the Markov Chains theory to calculate new multi-target spectral moments to fit a mt-QSAR model for 500 drugs tested in the literature against 16 parasite species and other 207 drugs no tested in the literature. The data were processed by LDA, classifying drugs as active or non-active against the different tested parasite species. The model correctly classified 311 out of 358 active compounds (86.9 %) and 2,328 out of 2,577 non-active compounds (90.3 %) in training series. Overall training performance was 89.9 %. Validation of the model was carried out by means of external predicting series. In these series the model classified correctly 157 out of 190 (82.6 %) antiparasitic compounds and 1,151 out of 1,277 non-active compounds (90.1 %). Overall predictability performance was 89.2 %. In addition four types of non Linear Artificial neural networks (ANN) were developed and compared with the mt-QSAR model. The improved ANN model had an overall training performance of 87 %. The equation of the model is the following:

$$Actv = 1.49 \cdot {}^{1}\mu_s(C_{uns}) + 1.12 \cdot {}^{5}\mu_s(C_{uns}) + 1.92 \cdot {}^{3}\mu_s(C_{sat}) + 0.53 \cdot {}^{4}\mu_s(X) + 1.71 \cdot {}^{1}\mu_s(\text{H-Het}) - 0.97 \cdot {}^{2}\mu_s(\text{H-Het}) - 5.21$$

$$\lambda = 0.52 \quad X^2 = 1904.6; \quad p < 0.001$$

The coefficient λ is the Wilk's statistic; statistic for the overall discrimination, χ^2 is the chi-square, and p the error level. In this equation, ${}^{k}\mu_s$ was calculated for the total (T) of atoms in the molecule or for specific collections of atoms. These collections are atoms with a common characteristic: H-Het: hydrogen bound to heteroatom, C_{uns}: unsaturated carbon atoms, C_{sat}: saturated Carbon atoms, X: halogen atoms.

4.4. mt-QSAR for Anti-Fungi Drugs

González-Díaz *et al.* [95] developed a unified Markov model to describe with a single linear equation the biological activity of 74 drugs tested in the literature against some of the fungi species selected from a list of 87 species (491 cases in total). The data were processed by LDA, classifying drugs as active or non-active against the different tested fungi species. The model correctly classified 338 out of 368 active compounds (91.85 %) and 89 out of 123 non-active compounds (72.36 %). Overall training predictability was 86.97 % (427 out of 491 compounds). Validation of the model was carried out by means of leave-species-out (LSO) procedure. After elimination step-by-step of all drugs tested against one specific species, the authors recorded the percentage of good classification of leave-out compounds (LSO-predictability). In addition, robustness of the model to the elimination of the compounds (LSO-robustness) was considered. This aspect was considered as the variation of the percentage of good classification of the modified model (Δ) in LSO with respect to the original one. Average LSO-predictability was 86.41 ± 0.95 % (average ± SD) and Δ = -0.55 %, being 6 the average number of drugs tested against each fungi species. Results for some of the 87 studied species were *Candida albicans*: 43 tested compounds, 100 % of LSO-predictability, Δ = -3.49 %; *Candida parapsilosis* 23, 100 %, Δ = -0.86 %; *Aspergillus fumigatus* 21, 95.20 %, Δ = 0.05 %; *Microsporum canis* 12, 91.60 %, Δ = -2.84 %; *Trichophyton mentagrophytes* 11, 100 %, Δ = -0.51 %; *Cryptococcus neoformans* 10, 90 %, Δ = -0.90 %. The equation of the model is the following:

$$Actv = -2.88 \cdot {}^{0}C_s(X) + 1.26 \cdot {}^{5}C_s(X) - 1.01 \cdot {}^{0}C_s(T) - 0.78 \cdot {}^{0}C_s(C_{uns}) + 0.94 \cdot {}^{3}C_s(X) - 0.76 \cdot {}^{4}C_s(T) - 1.17$$

$$\lambda = 0.53; \quad F(6,484) = 71.93; \quad p < 0.001$$

Where, λ is the Wilk's statistics, statistic for the overall discrimination, F is the Fisher ratio, and p is the error level. In this equation, ${}^{k}C_s$ were calculated for the totality (T) of the atoms in the molecule or for specific collections of atoms. These collections are atoms with a common characteristic: X: halogens and C_{uns}: unsaturated carbon atoms.

González-Díaz & Prado-Prado [87] selected pairs of antifungal drugs with similar/dissimilar species predicted-activity profile and represented them as a large network, They developed a mt-QSAR classification model, in which the outputs were the inputs of the aforementioned network. Overall model classification accuracy was 87.0 % (161 out of 185 compounds) in training, 83.4 % (50 out of 61) in validation, and 83.7 % for 288 additional antifungal compounds used to extend model validation for network construction. The network predicted has 59 nodes (compounds), 648 edges (pairs of compounds with similar activity), low coverage density d = 37.8 %, and distribution closer to normal than to exponential. The equation of the model is the following:

$$Actv = -0.49 \cdot {}^{A}\pi_5(s, C_{sp\&sp2}) - 2.57 \cdot {}^{A}\pi_0(s,X) + 1.43 \cdot {}^{A}\pi_0(s,\text{H-Het}) + 0.90$$

$$R_c = 0.75 \quad \lambda = 0,44 \quad p < 0.001$$

Where, Rc it is the canonical correlation coefficient, λ it is the Wilk's statistics; and p the error level. In this equation, the absolute probabilities ${}^{A}\pi_k$ calculated refers to:

1. ${}^{A}\pi_5(s, C_{sp\&sp2})$ all unsaturated carbon atoms (sp and sp2 atoms) and all atoms placed at five or least atoms from them.
2. ${}^{A}\pi_0(s,X)$ all halogens atoms.
3. ${}^{A}\pi_0(s,$ H-Het) all hydrogen atoms bound to a heteroatom (N, O, or S).

Prado-Prado *et al.* [96] used the Markov Chain theory to calculate new multi-target spectral moments to fit a mt-QSAR model that predicts the antifungal activity of more than 280 drugs against 90 fungi species. LDA was used to classify drugs as active or non-active against the different fungal species. The model correctly classified 12,434 out of 12,566 non-active compounds (98.95 %) and 421 out of 468

Table 2. Overall Train Accuracy, CV Predictability, and Models for Different Drug Side Effects

Side Effects	Train	CV	Model
			Breathing Manifestations
Infiltrated lung (IL)	92.3	98.1	$IL = -87.38+35.17\cdot{}^{0}\Gamma_M$
Bronchospasm (Brch)	100	100	$Brch = -176.69+50.12\cdot{}^{0}\Gamma_M$
Total	97.1	97.5	N= 35; U= 0.114; F= 255.7
			Cardiovascular manifestations
Exacerbations of angina pectoris (EAP)	100	100	$EAP = -112.70+59.90\cdot{}^{0}\Gamma_M-1.421\cdot{}^{4}\Gamma_M+0.689\cdot{}^{1}\Gamma_M$
Arrhythmias (Arr)	100	100	$Arr = -1,862.78+245.41\cdot{}^{0}\Gamma_M-5.67\cdot{}^{4}\Gamma_M+2.74\cdot{}^{1}\Gamma_M$
Edema or liquid retention in heart inadequacy (ELRHI)	100	98.9	$ELRHI = -594.65+138.47\cdot{}^{0}\Gamma_M-3.21\cdot{}^{4}\Gamma_M+1.56\cdot{}^{1}\Gamma_M$
Hypertension (HyperT)	100	100	$HyperT = -137.72+66.37\cdot{}^{0}\Gamma_M-1.60\cdot{}^{4}\Gamma_M+0.77\cdot{}^{1}\Gamma_M$
Hypotension (HypoT)	86.4	96.6	$HypoT = -411.95+115.20\cdot{}^{0}\Gamma_M-2.55\cdot{}^{4}\Gamma_M+1.23\cdot{}^{1}\Gamma_M$
Thromboembolism (Thr)	100	100	$Thr = -315.70+100.72\cdot{}^{0}\Gamma_M-2.45\cdot{}^{4}\Gamma_M+1.19\cdot{}^{1}\Gamma_M$
Total	97.6	96.8	N= 125; U= 0.003; F= 158.6
			Hematological manifestations
Agranulocytosis (Agr)	85	98.8	$Agr = -391.15+149.73\cdot{}^{0}\Gamma_M-12.99\cdot{}^{2}\Gamma_M+9.39\cdot{}^{1}\Gamma_M$
Hemolytic anemia (HA)	100	100	$HA = -903.99+227.83\cdot{}^{0}\Gamma_M-19.63\cdot{}^{2}\Gamma_M+14.20\cdot{}^{1}\Gamma_M$
Hemolytic anemia (in deficit of G6PD) (HAG6PD)	100	100	$HAG6PD = -162.71+96.29\cdot{}^{0}\Gamma_M-8.28\cdot{}^{2}\Gamma_M+5.98\cdot{}^{1}\Gamma_M$
Pancytopenia (Pan)	90	100	$Pan = -334.62+138.43\cdot{}^{0}\Gamma_M-11.85\cdot{}^{2}\Gamma_M+8.57\cdot{}^{1}\Gamma_M$
Platelet aggregation alterations (PAA)	100	1080	$PAA = -640.64+191.50\cdot{}^{0}\Gamma_M-16.22\cdot{}^{2}\Gamma_M+11.74\cdot{}^{1}\Gamma_M$
Total	95.1	99.1	N= 103; U= 0.012; F= 90.5
			Gastrointestinal manifestations
Constipation or ileo (Col)	90.9	100	$Col = -41.91+20.78\cdot{}^{0}\Gamma_M+0.33\cdot{}^{2}\Gamma_M-0.39\cdot{}^{3}\Gamma_M$
Diarrhea or colitis (DoC)	100	100	$DoC = -208.83+47.26\cdot{}^{0}\Gamma_M+0.60\cdot{}^{2}\Gamma_M-0.70\cdot{}^{3}\Gamma_M$
Diffuse hepatocellular Damage (DHD)	95	98.8	$DHD = -93.19+31.41\cdot{}^{0}\Gamma_M+0.63\cdot{}^{2}\Gamma_M-0.73\cdot{}^{3}\Gamma_M$
Mouth dryness (MD)	86.7	100	$MD = -65.40+26.19\cdot{}^{0}\Gamma_M+0.62\cdot{}^{2}\Gamma_M-0.72\cdot{}^{3}\Gamma_M$
Nausea or vomit (NoV)	100	100	$NoV = -1562.42+129.82\cdot{}^{0}\Gamma_M+2.52\cdot{}^{2}\Gamma_M-2.95\cdot{}^{3}\Gamma_M$
Pancreatitis (Pat)	100	100	$Pat = -53.90+23.70\cdot{}^{0}\Gamma_M+0.46\cdot{}^{2}\Gamma_M-0.53\cdot{}^{3}\Gamma_M$
Peptic or hemorragic ulceration (PoHU)	93.8	100	$PoHU = -63.21+25.70\cdot{}^{0}\Gamma_M+0.20\cdot{}^{2}\Gamma_M-0.24\cdot{}^{3}\Gamma_M$
Total	97	99.4	N= 164; U= 0.002; F= 190.9
			Dermal manifestations
Acne (Ac)	50	90	$Ac = -212.11+117.62\cdot{}^{0}\Gamma_M$
Alopecia (Alp)	100	100	$Alp = -1,456.17+309.04\cdot{}^{0}\Gamma_M$
Diverse erythema (DE)	91.7	100	$DE = -196.81+113.27\cdot{}^{0}\Gamma_M$
Photodermatitis (PhD)	100	100	$PhD = -268.05+123.31\cdot{}^{0}\Gamma_M$
Total			N= 50; U= 0.004; F= 3920
			Systemic phenomena
Anaphylaxis (Anph)	100	100	$Anph = -46.36+17.29\cdot{}^{0}\Gamma_M$
Lupus Erythematosus (LE)	100	100	$LE = -14.43+9.39\cdot{}^{0}\Gamma_M$
Fever (Fv)	100	100	$Fv = -133.28+29.55\cdot{}^{0}\Gamma_M$
Total	100	100	N= 47; U= 0.042; F= 506.1
			Endocrine manifestations
Galactorrhea (amenorrhea) (Gal)	100	100	$Gal = -71.25+41.25\cdot{}^{0}\Gamma_M$
Livid decrease and impotence (LDI)	100	100	$LDI = -98.16+48.52\cdot{}^{0}\Gamma_M$
Thyroid function test disorders (TFTD)	88.9	94.4	$TFTD = -30.62+26.76\cdot{}^{0}\Gamma_M$
Total	97	95	N= 33; U= 0.12; F= 110
			Metabolic manifestations
Hyperglicemia (HyperG)	100	100	$HyperG = -56.91+48.06\cdot{}^{0}\Gamma_M$
Hypopotassemia (HypoP)	100	100	$HypoP = -141.22+75.92\cdot{}^{0}\Gamma_M$

Table 2. cont....

Side Effects	Train	CV	Model
			Breathing Manifestations
Total	100	100	N=18; U= 0.086; F= 170.8
			Neurological manifestations
Convulsions (Cvs)	100	100	$Cvs = -597.82 + 166.92\,{}^{0}\Gamma_M$
Extrapyramidal effects (EE)	100	100	$EE = -252.34 + 108.36\,{}^{0}\Gamma_M$
Total	100	100	N= 33; U= 0.026; F= 11.58
			Psychiatric manifestations
Deliriums or confusional states (DoCS)	100	100	$DoCS = -211.26 + 70.9\,{}^{0}\Gamma_M + 0.79\,{}^{1}\Gamma_M - 1.46\,{}^{5}\Gamma_M$
Dysfunctions of the dream (DD)	100	100	$DD = -84.02 + 44.48\,{}^{0}\Gamma_M + 0.28\,{}^{1}\Gamma_M - 0.46\,{}^{5}\Gamma_M$
Somnolence (Snl)	90.9	96.6	$Snl = -270.26 + 80.03\,{}^{0}\Gamma_M + 0.41\,{}^{1}\Gamma_M - 0.60\,{}^{5}\Gamma_M$
Total	96.4	94.6	N= 55; U= 0.04; F= 66.5
			Muscular-skeletal manifestations
Myopathy or myalgia (MoM)	100	100	$MoM = -276.33 + 92.77\,{}^{0}\Gamma_M - 0.49\,{}^{5}\Gamma_M$
Osteoporosis (Ost)	100	100	$Ost = -61.14 + 43.41\,{}^{0}\Gamma_M - 0.23\,{}^{5}\Gamma_M$
Total	100	100	N= 23; U= 0.027; F= 361.4

active compounds (89.96 %). Overall training predictability was 98.63 %. Validation of the model was carried out by means of external predicting series, classifying 6,216 out of 6,277 non-active compounds and 215 out of 239 active compounds. Overall training predictability was 98.7 %. The equation of the model is the following:

$$Actv = -3.44 \cdot {}^{5}\mu_s(\text{Het}) - 3.18 \cdot {}^{2}\mu_s(\text{H-Het}) - 3.85 \cdot {}^{3}\mu_s(C_{sat}) + 4.76 \cdot {}^{4}\mu_s(C_{sat}) - 4.61 \cdot {}^{5}\mu_s(C_{sat}) + 28.26 \cdot {}^{0}\mu_s(\text{T}) - 29.26$$

$$\lambda = 0.33; \quad X^2 = 14367.94; \quad p < 0.001$$

Where, χ^2 is the Chi-square, and p the error level. In this equation, ${}^{k}\mu_s$ were calculated for the total (T) of atoms in the molecule or for specific collections of atoms. These collections are atoms with a common characteristic: Het: heteroatom, H-Het: hydrogen bound to heteroatom, C_{sat}: saturated carbon atoms.

4.5. mt-QSTR for Drug Side Effects

González-Díaz *et al.* [97] developed a general Markov model that describes 39 different drug side effects grouped in 11 affected systems for 301 drugs, being 686 cases finally. The data was processed by LDA, classifying drugs according to their specific side effects. The average percentage of good classification and number of compounds used in the training/predicting sets were 100/100 % for systemic phenomena (47 out of 47)/(12 out of 12) and metabolic (18 out of 18)/(5 out of 5), muscular–skeletal (23 out of 23)/(6 out of 6) and neurological manifestations (33 out of 33)/(8 out of 8); 97.6/96.7 % for cardiovascular manifestation (122 out of 125)/(30 out of 31); 97.1/97.5 % for breathing manifestations (34 out of 35)/(8 out of 9); 97/99.4 % for gastrointestinal manifestations (159 out of 164)/(40 out of 41); 97/95 % for endocrine manifestations (32 out of 33)/(7 out of 8); 96.4/94.6 % for psychiatric manifestations (53 out of 55)/(13 out of 14); 95.1/99.1 % for hematological manifestations (98 out of 103)/(25 out of 26) and 88/92.3 % for dermal manifestations (44 out of 50)/(12 out of 13). This article develops a model that encompasses a large number of side effects grouped in specific organ systems in a single stochastic framework for the first time. The models developed for the different drug side effects are shown in the Table **2**.

Cruz-Monteagudo & González-Díaz [98] developed a general Markov model that describes 21 different drug side effects grouped in 10 affected biological systems for 193 drugs. The data were processed by LDA, classifying drugs according to their specific side effects. The average percentage of good classification and number of compounds used in the training/predicting sets were 92.6/91.7 % for cardiovascular manifestation (25 out of 27)/(18 out of 20); 89.3/83.9 % for dermal manifestations (25 out of 18)/(18 out of 21); 88.9/88.9 % for endocrine manifestations (16 out of 18)/(12 out of 14); 88.9/88.2 % for psychiatric manifestations (32 out of 36)/(24 out of 27); 88.5/85.6 % for systemic phenomena (23 out of 26)/(17 out of 20); 85.7/91.7 % for gastrointestinal manifestations (36 out of 42)/(29 out of 32); 83.3/79.2 % for metabolic manifestations (15 out of 18)/(11 out of 14); 81.8/78.0 % for neurological manifestations (27 out of 33)/(20 out of 25); 75.0/74.0 % for hematological manifestations (36 out of 48)/(27 out of 36) and 74.3/72.8 % for breathing manifestations (26 out of 35)/(19 out of 26). Application of back-projection analysis provides physic interpretation in structural terms through molecular graphics of the toxic effects predicted with these QSTR models. This article develops a mathematical model that encompasses a large number of drugs side effects grouped in specifics systems using stochastic entropies of interaction ($\Theta_k(j)$) for the first time. The models developed for the different drug side effects are shown in the Table **3**.

Cruz-Monteagudo *et al.* [99] developed a general Markov model that describes 19 different drug side effects grouped in eight affected biological systems for 178 drugs, being 270 cases finally. The data were processed by LDA, classifying drugs according to their specific side effects. The average percentage of good classification and number of compounds used in the training/predicting sets

Table 3. Overall train Accuracy, CV Predictability, and Models for Different Drug Side Effects

Side Effects	Train	CV	Model
			Breathing Manifestations
Infiltrated lung (IL)	70.0	61.5	IL= −14.21+3.66·Θ_1(C_{uns}) − 3.61·Θ_5(C_{uns}) − 1.76·Θ_5(C_{sat})+16.03·Θ_1(T)
Bronchospasm (Brch)	77.3	79.5	Brch= −16.43+10.8166·Θ_1 (C_{uns}) − 9.58·Θ_5(C_{uns}) − 0.12·Θ_5(C_{sat})+13.18·Θ_1(T)
Total	74.3	72.8	N= 35; U= 0.723; F= 2.88; p= 0.0396
			Cardiovascular manifestations
Hypertension (HyperT)	84.6	82.7	HyperT= −36.85 − 1.65·Θ_2(C_{uns}) + 229.40·Θ_5(Het) + 45.66·Θ_0(T) − 225.57·Θ_4(Het)
Thromboembolism (Thr)	100.0	100.0	Thr= −52.17−4.18·Θ_2 (C_{uns})+344.85·Θ_5(Het) +56.58·Θ_0(T) −337.91·Θ_4(Het)
Total	92.6	91.7	N= 27; U= 0.317; F= 11.87; p= 0.0000
			Hematological manifestations
Agranulocytosis (Agr)	75.0	76.3	Agr= −58.04+8.94·Θ_0(HX) −22.44·Θ_5 (C_{uns}) +48.18·Θ_5(T)+33.10·Θ_0(C_{uns})
Hemolytic anemia (HA)	75.0	72.3	HA= − 69.03 + 11.07·Θ_0($_{HX}$) − 23.17·Θ_5 (C_{uns}) + 51.56·Θ_5(T) + 34.87·Θ_0(C_{uns})
Total	75.0	74.0	N= 48; U= 0.670; F= 5.29; p= 0.0015
			Gastrointestinal manifestations
Constipation or ileo (Col)	90.9	95.5	Col= −6.81+2.34·Θ_2(C_{uns}) +1.31·Θ_5 (Het) −1.00·Θ_5 (HX)
Pancreatitis (Pat)	93.3	93.3	Pat= −8.08+1.28 ·Θ_2(C_{uns}) +3.22·Θ_5 (Het)+1.86·Θ_5 (HX)
Peptic or hemorrhagic ulceration (PoHU)	75.0	87.5	PoHU= −13.78+3.05·Θ_2(C_{uns}) +3.23·Θ_5 (Het) −0.67·Θ_5 (HX)
Total	85.7	91.7	N= 42; U= 0.252; F= 12.26; p= 0.0000
			Dermal manifestations
Alopecia (Alp)	87.5	82.2	Alp= −23.05+7.63·Θ_0(C_{sat}) +6.44·Θ_1(C_{uns}) +30.43·Θ_0 (Hal)
Photodermatitis (PhD)	91.7	85.4	PhD= −28.09+6.71 ·Θ_0(C_{sat}) +8.04·Θ_1 (C_{uns}) +38.29·Θ_0 (Hal)
Total	89.3	83.9	N= 28; U= 0.484; F= 8.54; p= 0.0005
			Systemic phenomena
Anaphylaxis (Anph)	85.7	83.9	Anph= −95.40+103.44 ·Θ_0(T) +5.19·Θ_0 (Het) +0.93·Θ_5 (HX)
Lupus Erythematosus (LE)	91.7	87.5	LE= −73.78+92.28·Θ_0(T) +2.90·Θ_0 (Het) +1.97·Θ_5 (HX)
Total	88.5	85.6	N= 26; U= 0.437; F= 9.45; p= 0.0003
			Endocrine manifestations
Galactorrhea (amenorrhea) (Gal)	88.9	88.9	Gal= −3.46+18.78·Θ_5(C_{sat}) − 17.18·Θ_4 (C_{sat})
Thyroid function test disorders (TFTD)	88.9	88.9	TFTD= −1.94 − 76.55·Θ_5(C_{sat})+77.27·Θ_4 (C_{sat})
Total	88.9	88.9	N= 18; U= 0.529; F= 6.68; p= 0.0084
			Metabolic manifestations
Hyperglicemia (HyperG)	77.8	88.9	HyperG= −12.00+55.68·Θ_5(HX) − 55.57·Θ_3(HX) +13.02·Θ_1 (T)
Hypopotassemia (HypoP)	88.9	69.4	HypoP= −16.93+126.13·Θ_5(HX) −124.54·Θ_3(HX) +15.04·Θ_1 (T)
Total	83.3	79.2	N= 18; U= 0.586; F= 3.29; p= 0.0521
			Neurological manifestations
Convulsions (Cvs)	90.0	82.5	Cvs= −3.68+0.21·Θ_3(C_{uns})+1.86·Θ_0(C_{uns})+21.78·Θ_5 (Hal) +3.95·Θ_0(Hal) −25.24·Θ_2(Hal)
Extrapyramidal effects (EE)	70.0	71.1	EE= −5.98+2.09·Θ_5(C_{uns})+0.09·Θ_0(C_{uns})+6.98·Θ_5 (Hal) −430.00·Θ_0(Hal)+342.48·Θ_2(Hal)

Table 3. cont...

Side Effects	Train	CV	Model
			Breathing Manifestations
Total	81.8	78.0	N= 33; U= 0.614; F= 3.40; p= 0.0164
			Psychiatric manifestations
Dysfunctions of the dream (DD)	85.7	85.7	$DD = -4.88 + 2.58 \cdot \Theta_5(C_{uns})$
Somnolence (Snl)	90.9	89.8	$Snl = -13.03 + 4.43 \cdot \Theta_5(C_{uns})$
Total	88.9	88.2	N= 36; U= 0.480; F= 36.81; p= 0.0000

$\Theta_k(1)$ represents a global molecular index. $\Theta_k(C_{sat})$, $\Theta_k(C_{uns})$, $\Theta_k(Hal)$, $\Theta_k(Het)$ and $\Theta_k(HX)$ represent local molecular indices describing saturated carbon atom, unsaturated carbon atom, halogens, heteroatoms and hydrogen bonded to heteroatoms, respectively

Table 4. Overall Train Accuracy, CV Predictability, and Models for Different Drug Side Effects

Side Effects	Train	CV	Model
			Breathing Manifestations
Infiltrated lung (IL)	84.6	72.7	$IL = -41.28 + 44.95 \cdot {}^{A}\pi_1(C_{uns}) + 316.99 \cdot {}^{A}\pi_0(C_{sat}) + 85.53 \cdot {}^{A}\pi_5(C_{uns}) + 32.62 \cdot {}^{A}\pi_1(Hal)$
Bronchospasm (Brch)	72.7	78.4	$Brch = -48.18 + 102.39 \cdot {}^{A}\pi_1(C_{uns}) + 341.35 \cdot {}^{A}\pi_0(C_{sat}) + 41.86 \cdot {}^{A}\pi_5(C_{uns}) + 16.99 \cdot {}^{A}\pi_1(Hal)$
Total	77.1	75.7	N= 35; U= 0.705; F= 3.13; p= 0.0289
			Cardiovascular manifestations
Exacerbations of angina pectoris (EAP)	80.0	77.5	$EAP = -7.88 + 21.91 \cdot {}^{A}\pi_4(C_{uns}) + 21.39 \cdot {}^{A}\pi_5(Het) + 17.52 \cdot {}^{A}\pi_3(HX) - 31.82 \cdot {}^{A}\pi_5(Hal)$
Hypertension (HyperT)	76.9	76.9	$HyperT = -6.57 + 24.57 \cdot {}^{A}\pi_4(C_{uns}) + 3.46 \cdot {}^{A}\pi_5(Het) + 20.33 \cdot {}^{A}\pi_3(HX) + 4.33 \cdot {}^{A}\pi_5(Hal)$
Thromboembolism (Thr)	78.6	98.2	$Thr = -4.73 + 7.34 \cdot {}^{A}\pi_4(C_{uns}) + 31.02 \cdot {}^{A}\pi_5(Het) - 1.79 \cdot {}^{A}\pi_3(HX) - 13.36 \cdot {}^{A}\pi_5(Hal)$
Total	78.4	85.1	N= 37; U= 0.407; F= 4.39; p= 0.0003
			Gastrointestinal manifestations
Constipation or ileo (CoI)	90.9	97.7	$CoI = -4.78 + 19.91 \cdot {}^{A}\pi_2(C_{uns}) + 8.86 \cdot {}^{A}\pi_1(Het) - 9.49 \cdot {}^{A}\pi_5(HX)$
Pancreatitis (Pat)	93.3	93.3	$Pat = -6.68 + 8.99 \cdot {}^{A}\pi_2(C_{uns}) + 32.13 \cdot {}^{A}\pi_1(Het) + 24.31 \cdot {}^{A}\pi_5(HX)$
Peptic or hemorrhagic ulceration (PoHU)	87.5	87.5	$PoHU = -10.96 + 27.29 \cdot {}^{A}\pi_2(C_{uns}) + 29.54 \cdot {}^{A}\pi_1(Het) - 5.35 \cdot {}^{A}\pi_5(HX)$
Total	90.5	92.3	N=42; U= 0.259; F=11.88; p= 0.0000
			Dermal manifestations
Acne (Ac)	80.0	82.5	$Ac = -5.24 + 27.41 \cdot {}^{A}\pi_2(C_{uns}) - 46.53 \cdot {}^{A}\pi_5(Hal) + 0.15 \cdot {}^{A}\pi_1(Het) + 7.83 \cdot {}^{A}\pi_3(HX) + 178.39 \cdot {}^{A}\pi_0(Hal)$
Alopecia (Alp)	75.0	89.1	$Alp = -6.81 + 16.59 \cdot {}^{A}\pi_2(C_{uns}) - 291.85 \cdot {}^{A}\pi_5(Hal) + 43.82 \cdot {}^{A}\pi_1(Het) - 22.73 \cdot {}^{A}\pi_2(HX) + 361.79 \cdot {}^{A}\pi_0(Hal)$
Photodermatitis (PhD)	91.7	85.4	$PhD = -11.48 + 37.57 \cdot {}^{A}\pi_2(C_{uns}) - 233.34 \cdot {}^{A}\pi_5(Hal) + 25.66 \cdot {}^{A}\pi_1(Het) - 7.21 \cdot {}^{A}\pi_2(HX) + 399.68 \cdot {}^{A}\pi_0(Hal)$
Total	81.6	86.2	N= 38; U= 0.347; F=4.32; p= 0.0001
			Systemic phenomena
Anaphylaxis (Anph)	85.7	82.1	$Anph = -16.08 + 97.23 \cdot {}^{A}\pi_0(Het) - 14.83 \cdot {}^{A}\pi_5(HX) + 217.40 \cdot {}^{A}\pi_0(C_{sat}) - 100.21 \cdot {}^{A}\pi_5(C_{sat})$
Lupus Erythematosus (LE)	91.7	91.7	$LE = -7.67 + 57.99 \cdot {}^{A}\pi_0(Het) + 3.69 \cdot {}^{A}\pi_5(HX) + 111.03 \cdot {}^{A}\pi_0(C_{sat}) - 44.93 \cdot {}^{A}\pi_5(C_{sat})$
Total	88.5	86.5	N= 26; U= 0.404; F= 7.76; p= 0.0005
			Endocrine manifestations
Galactorrhea (amenorrhea) (Gal)	100.0	94.4	$Gal = -8.60 + 4{,}767.30 \cdot {}^{A}\pi_5(C_{sat}) - 11{,}109.50 \cdot {}^{A}\pi_4(C_{sat}) + 7{,}361.80 \cdot {}^{A}\pi_3(C_{sat}) - 966.90 \cdot {}^{A}\pi_1(C_{sat})$

Table 4. cont...

Side Effects	Train	CV	Model
			Breathing Manifestations
Thyroid function test disorders (TFTD)	100.0	97.2	$TFTD = -2.71 + 745.71 \cdot {}^{A}\pi_5(C_{sat}) - 2{,}972.21 \cdot {}^{A}\pi_4(C_{sat}) + 2{,}630.04 \cdot {}^{A}\pi_3(C_{sat}) - 384.09 \cdot {}^{A}\pi_1(C_{sat})$
Total	100	95.8	N= 18; U= 0.309; F= 7.28; p= 0.0026
			Neurological manifestations
Convulsions (Cvs)	90.0	87.5	$Cvs = -3.62 + 735.92 \cdot {}^{A}\pi_5(Hal) + 598.03 \cdot {}^{A}\pi_0(Hal) - 1{,}163.01 \cdot {}^{A}\pi_2(Hal) - 51.39 \cdot {}^{A}\pi_5(C_{uns}) + 69.96 \cdot {}^{A}\pi_2(C_{uns})$
Extrapyramidal effects (EE)	70.0	61.5	$EE = -5.38 - 138.26 \cdot {}^{A}\pi_5(Hal) - 4{,}193.30 \cdot {}^{A}\pi_0(Hal) + 3{,}322.32 \cdot {}^{A}\pi_2(Hal) + 18.90 \cdot {}^{A}\pi_3(C_{uns}) + 3.88 \cdot {}^{A}\pi_2(C_{uns})$
Total	81.8	77.3	N= 33, U= 0.632; F= 3.14; p= 0.0231
			Psychiatric manifestations
Deliriums or confusional states (DoCS)	73.7	69.7	$DoCS = -21.07 + 76.99 \cdot {}^{A}\pi_2(C_{uns}) + 75.11 \cdot {}^{A}\pi_5(C_{sat})$
Somnolence (Snl)	77.3	77.0	$Snl = -28.82 + 92.94 \cdot {}^{A}\pi_2(C_{uns}) + 83.05 \cdot {}^{A}\pi_5(C_{sat})$
Total	75.6	75.0	N= 41; U= 0.660; F= 9.79; p= 0.0004

$\pi_5(C_{sat})$, $\pi_4(C_{uns})$, $\pi_3(Hal)$, $\pi_2(Het)$ and $\pi_1(HX)$ represent molecular indices describing saturated carbon atom, unsaturated carbon atom, halogens, heteroatoms and hydrogen bonded to heteroatoms, respectively.

were 100/95.8% for endocrine manifestations, (18 out of 18)/(13 out of 14); 90.5/92.3 % for gastrointestinal manifestations, (38 out of 42)/(30 out of 32); 88.5/86.5 % for systemic phenomena, (23 out of 26)/(17 out of 20); 81.8/77.3 % for neurological manifestations, (27 out of 33)/(19 out of 25); 81.6/86.2 % for dermal manifestations, (31 out of 38)/(25 out of 29); 78.4/85.1 % for cardiovascular manifestation, (29 out of 37)/(24 out of 28); 77.1/75.7 % for breathing manifestations, (27 out of 35)/(20 out of 26) and 75.6/75 % for psychiatric manifestations, (31 out of 41)/(23 out of 31). Additionally a back-projection analysis was carried out for two ulcerogenic drugs to prove in structural terms the physical interpretation of the models obtained. This article develops a mathematical model that encompasses a large number of drugs side effects grouped in specifics biological systems using stochastic absolute probabilities of interaction (${}^{A}\pi_k(j)$) for the first time. The models developed for the different drug side effects are shown in the Table **4**.

4.6. mt-QSPR for Drug Distribution

Predicting tissue and environmental distribution of chemicals is of major importance for environmental and life sciences. Most of the molecular descriptors used in computational prediction of chemicals partition behavior consider molecular structure but ignore the nature of the partition system. Consequently, computational models derived up-to-date are restricted to the specific system under study. Here, a free energy-based descriptor (ΔG_k) is introduced, which circumvent this problem. Based on ΔG_k, Cruz-Monteagudo & González-Díaz [100] developed for the first time a single linear classification model to predict the partition behavior of a broad number of structurally diverse drugs and other chemicals (1,300) for 38 different partition systems of biological and environmental significance. The model presented training/predicting set accuracies of 91.79/88.92 %. In addition, inversion of the partition direction for each one of the 38 partition systems evidenced that the model correctly classified 89.08 % of compounds. Other 10 different classification models (linear, neural networks, and genetic algorithms) were also tested for the same purposes. None of these computational models was better than the proposed model, indicating that this approach capture the main aspects that govern chemicals partition in different systems. The equation of the model is the following:

$$PL = 3.1 \cdot \Delta G_5(T) - 1.01 \cdot \Delta G_5(HX) + 1.54 \cdot \Delta G_0(T) - 0.89 \Delta G_1(C_{sat}) + 0.36$$

Where ΔG_k is the free energy- based descriptor for a given partition system after k steps. It has been calculated for the total (T) of atoms in the molecule or for specific collections of atoms. These collections are atoms with a common characteristic: HX: hydrogen bound to heteroatom, C_{sat}: saturated carbon atoms.

In a later paper, González-Díaz *et al.* [101] defined the multi-system partition Complex Networks (MSP-CNs) as large graphs composed by nodes (chemicals) interconnected by arcs if a pair of chemicals have similar partition in a given system. They built the first MSP-CN based on a multi-target QSPR (mt-QSPR). The model is based on the spectral moments (π_k) of a molecular Markov matrix weighted with atomic parameters that depend on both the nature of the atom and the partition system. The mt-QSPR predicted 90.6% of 413 compound/system pairs in training series and 90.0% in validation. The MSP-CN predicted presents 413 nodes, 2,060 edges, average node degree 9.9, and only 7.7% drugs are unconnected. The model was used to study the biophysical phenomena of transport or distribution of G1 (a novel antimicrobial drug) to different rat tissues. Predicted probabilities (P) coincide with low experimental partition coefficients (logPC) reported herein for the first time in skin (P=0.455; logPC=−0.02b0→U), heart (0.453; −0.02→U), and brain (0.324; −0.34→U). The Kamada–Kawai algorithm evidenced the community structure of the MSP-CN and clus-

ters G1 into three different communities of the U-type drugs. These results coincide with the low distribution of G1 to these tissues and consequently have low expected drug side effect. The equation of the model is the following:

$$PL = 0.84 \cdot \pi_0(s_2 / s_1)_{Cuns} - 5.79 \cdot \pi_1(s_2 / s_1)_{H-Het} + 2.49 \cdot \pi_5(s_2 / s_1)_{Csat} - 7.21$$

Where, the terms C_{uns}, H-Het and C_{sat} express that they calculate local spectral moments by summing up only unsaturated carbon atoms, labile hydrogen or saturated carbon atoms rather than all the atoms present. The present LDA model showed a p-level <0.01, which means that accepting the model as valid presupposes an error level lower than 1% in the separation of L and U-type of compounds with respect to different release systems.

4.7. mt-QSAR for Drug-Target Pairs

There are many drugs described with very different affinity to a large number of receptors. Viña *et al.* [102] selected drug-receptor pairs (DRPs) of affinity/nonaffinity drugs to similar/dissimilar receptors and represented them as a large network, which may be used to identify drugs that can act on a receptor. In addition they developed a mt-QSAR classification model. Overall model classification accuracy was 72.25 % (1,390 of 1,924 compounds) in training and 72.28 % (459 of 635) in cross-validation. The equation of the model is the following:

$$\begin{aligned} S_{pred} &= 85.4 \cdot \alpha(d) - 70.5 \cdot \alpha(f_{0bp}, d) - 0.5 \cdot \alpha(f_{121aa}, d) - 15.0 \cdot \alpha(f_{181aa}, d) \\ &- 51.0 \cdot \log P(d) + 51.0 \cdot \log P(f_{61aa}, d) + 2.9 \cdot \chi(f_{241aa}, d) + 20.5 \cdot \eta(d) \\ &- 15.8 \cdot \eta(f_{0aa}, d) - 7.1 \cdot \eta(f_{181aa}, d) - 1.3 \end{aligned}$$

$$N = 1924; \quad {}^2\chi = 376.19; \quad p < 0.001$$

Where α_d is the polarizability, log P_d is the logarithm of the water/*n*-octanol partition coefficient, η is the hardness, χ_d is the molecular electronegativity, N is the number of cases (DRPs) used to train the model and chi-square (χ^2) is the statistic used to demonstrate that the model significantly discriminates between DRPs of compounds with affinity (aDRP) or nonaffinity (iDRP) to the receptor, at a $p < 0.001$ level of error. Following the notation given above, for example $\alpha(f_{0bp},d)=\alpha(d) - \alpha(f_{0bp})$ is the difference between the polarizability of the drug and the receptor region from 0 to 60bp.

Outputs of this mt-QSAR model were used as inputs to build a network. The observed network has 1,735 nodes (DRPs), 1,754 edges or pairs of DRPs with similar drug-target affinity (sPDRPs), and low coverage density $d = 0.12$ %. The predicted network has 1,735 DRPs, 1857 sPDRPs, and also low coverage density $d = 0.12$ %. After an edge-to-edge comparison (chi-square = 9420.3; $p < 0.005$), it has been demonstrated that the predicted network is significantly similar to the one observed and both have a distribution closer to exponential than to normal.

5. MOLECULAR NETWORKS

The aim of this section is to show how the network theory can be used in drug design and in the study of molecular and metabolic processes that are crucial for the survival of the organisms. The understanding of these processes is the first step to design strategies to treat diseases efficiently.

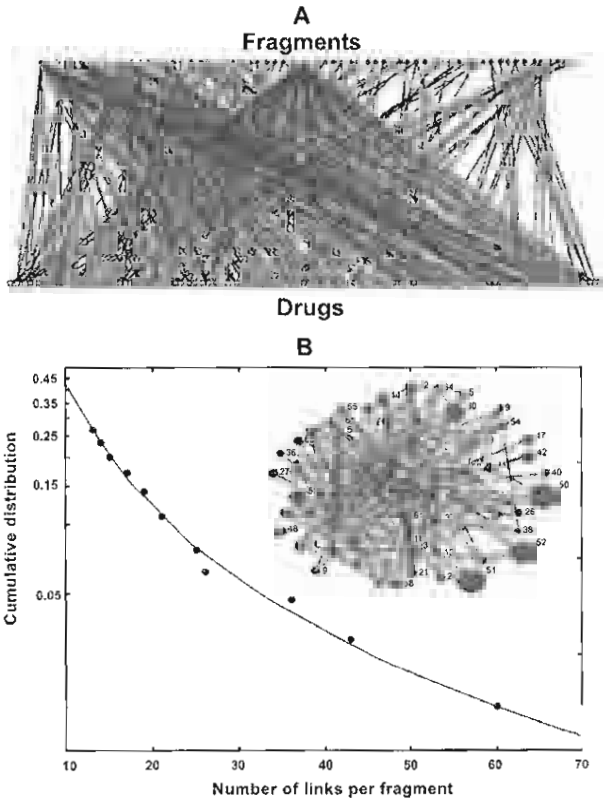


Fig. (6A). Bipartite network of fragments to drugs in which nodes at the top represent each of the 65 fragments found in this work to contribute to human serum albumin binding, and the nodes at the bottom represent each of the drugs in our data set. One fragment is connected to a drug if this fragment exists in the drug. (**B**) Linear-log plot of the cumulative distribution of the number of links per fragments in the fragment-fragment network. In this network, nodes represent the fragments found to contribute to human serum albumin binding, and two fragments are connected if they appear simultaneously in at least one drug. The upward curved line is characteristic of power-law (scale-free) distributions. Reproduced from [105].

5.1. Drug Fragment Networks

Fragment-based drug design is a tool for drug discovery that has emerged in the past decade. The starting point of this approach is always a chemical entity (typically MW 150-200), a fragment, with low affinity for the selected target. Fragments should satisfy key features such as diversity, reduced structural complexity, aqueous solubility and availability. Once selected, a fragment must undergo a heavy elaboration to improve binding affinity, at the same time acquiring drug-like properties. There are two main ways to go on at this point. The most common one is the so-called 'fragment evolution', consisting of a stepwise and systematic addition of chemical functionalities to the starting fragment core, together with a continuous feedback for pharmacological and physicochemical properties. The second one, less common but with great potential, is 'fragment linking': when two or more fragment hits are found to bind in adjacent regions of the target protein, they can be linked through appropriate spacers to rapidly produce a single molecule with much higher binding affinity [103, 104].

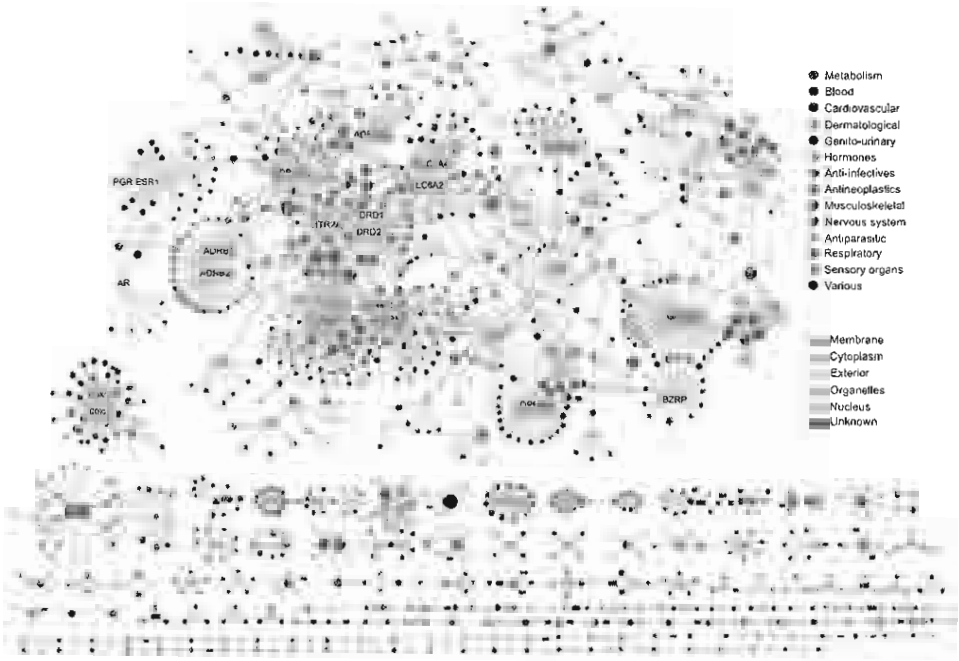


Fig. (7). Drug–target network (DT network). The DT network is generated by using the known associations between FDA-approved drugs and their target proteins. Circles and rectangles correspond to drugs and target proteins, respectively. A link is placed between a drug node and a target node if the protein is a known target of that drug. The area of the drug (protein) node is proportional to the number of targets that the drug has (the number of drugs targeting the protein). Color codes are given in the legend. Drug nodes (circles) are colored according to their Anatomical Therapeutic Chemical Classification, and the target proteins (rectangular boxes) are colored according to their cellular component obtained from the Gene Ontology database. Reproduced from [202].

Estrada *et al.* [105] in their *in silico* analysis of drug binding to human serum albumin built a bipartite network of fragments-to-molecules as illustrated in Fig. (**6**). Here, a link exists between a drug and a fragment only if the drug contains that fragment. Possible projections of this network give rise to a drug-drug or a fragment-fragment network. In the first case, all nodes represent drugs, and two drugs are connected if they share at least one fragment. In the second case, all nodes represent fragments, and two fragments are connected if they appear simultaneously in at least one drug. The authors studied the topology of the fragment-fragment network, observing that it displays small-world and scale-free characteristics.

Tanaka *et al.* [106] have shown that small-world phenomena are observed not only in existing chemical libraries but also in virtual libraries generated from structurally diverse fragments when represented as networks (nodes = chemical compounds; edges = similarities between chemical compounds). On the basis of this observation, they propose that an efficient compound-prioritization method of fragment-based drug discovery would be to select those fragments as a starting point such that the linked compounds become hubs in the library and therefore allow identification of many similar compounds when all-to-all fragment linkings are performed. Moreover, their analyses indicated that the variety of linkers had a marked influence on the network structure and thus on the diversity of the compounds synthesized by linking fragment hits.

5.2. Drug-Target Networks

The vast majority of successful drugs achieve their activity by binding to, and modifying the activity of, a protein [107]. Therefore, the search for interactions between compounds and proteins (targets) is an important part of drug discovery. Estimations of the total number of drug targets are presently dominated by analyses of the human genome, which are limited for various reasons, including the inability to infer the existence of splice variants or interactions between the encoded proteins from gene sequences alone, and the fact that the function of most of the DNA in the genome remains unclear [108]. Currently, target counts are of the order of 10^2, whereas estimations of the number of potential drug targets are an order of magnitude higher [107-109]. One interesting approach to the study of the drug-target interactions is the network theory. Yildirim *et al.* [202] built a bipartite graph composed of US Food and Drug Administration–approved drugs and proteins linked by drug–target binary associations (Fig. 7). The resulting network connects most drugs into a highly interlinked giant component, with strong local clustering of drugs of similar types according to Anatomical Therapeutic Chemical classification. Topological analyses of this network quantitatively showed an overabundance of 'follow-on' drugs, that is, drugs that target already targeted proteins. By including drugs currently under investigation, it was identified a trend toward more functionally diverse targets improving poly-pharmacology. To analyze the relationships between drug targets and disease-gene products, the authors measured the shortest distance between both sets of proteins in current models of the human interactome network. Significant differences in distance were found between etiological and palliative drugs.

Mestres *et al.* [110], using seven different databases, were able to assemble a total of 4,767 interactions between 802 drugs and 480 targets, which means that on average

every drug is currently acknowledged to interact with 6 targets. Their results confirm that the topology of drug-target networks depends implicitly on data completeness, drug properties, and target families. Csermely *et al.* [111] reviewed the efficiency of multi-target drugs from the point of view of the network theory, suggesting that partial inhibition of a small number of targets can be more efficient than the complete inhibition of a single target. Janga & Tzakos [112] formalize the definition of a drug-target network by decomposing it into drug, target and disease spaces and provide an overview of its structure and organizational principles. They discuss advances made in developing promiscuous drugs following the paradigm of poly-pharmacology and reveal their advantages over traditional drugs for targeting diseases such as cancer. They suggest that drug-target networks can be decomposed to be studied at a variety of levels and argue that such network-based approaches have important implications in understanding disease phenotypes and in accelerating drug discovery. They also discuss the potential and scope network pharmacology promises in harnessing the vast amount of data from high-throughput approaches for therapeutic advantage.

Targets for drugs have so far been predicted on the basis of molecular or cellular features, for example, by exploiting similarity in chemical structure or in activity across cell lines. Campillos *et al.* [113] have used phenotypic side-effect similarities to infer whether two drugs share a target. Applied to 746 marketed drugs, a network of 1,018 side effect–driven drug-drug relations became apparent, 261 of which are formed by chemically dissimilar drugs from different therapeutic indications. They experimentally tested 20 of these unexpected drug-drug relations and validated 13 implied drug-target relations by *in vitro* binding assays, of which 11 reveal inhibition constants equal to less than 10 μM. Nine of these were tested and confirmed in cell assays, documenting the feasibility of using phenotypic information to infer molecular interactions and hinting at new uses of marketed drugs. Yamanishi *et al.* [114] characterized four classes of drug–target interaction networks in humans involving enzymes, ion channels, G-protein-coupled receptors (GPCRs) and nuclear receptors, and revealed significant correlations between drug structure similarity, target sequence similarity and the drug–target interaction network topology. They developed new statistical methods to predict unknown drug–target interaction networks from chemical structure and genomic sequence information simultaneously on a large scale. The originality of the proposed method lies in the formalization of the drug–target interaction inference as a supervised learning problem for a bipartite graph, the lack of need for 3D structure information of the target proteins, and in the integration of chemical and genomic spaces into a unified space that they call 'pharmacological space'. In the results, they demonstrate the usefulness of the proposed method for the prediction of the four classes of drug–target interaction networks.

5.3. Protein Interaction Networks (PINs)

Protein-protein networks also known as Protein interaction networks (PINs) play an important role in understanding the functional and organizational principles of biological processes. Promising computational techniques for key systems biology research problems such as identification of signaling pathways, novel protein function prediction, and the study of disease mechanisms, are based on topological characteristics of the protein interactome [115]. A fully extended PIN can consist of upwards of several thousand nodes and edges. To simplify analysis, smaller child samples are often used in substitution of the global network. Blayney *et al.* [116] studied the impact of different levels of sampling on six PINs suggesting that restricting analysis to the first network level, using metrics such as degree and betweenness centrality, could lead to misrepresentative results, omitting potentially significant nodes. Fault-tolerance analysis also indicates that key nodes within the second network level, and above, contribute to the stability of the global network. Yook *et al.* [117] compared four available databases that approximate the PIN of the yeast, *Saccharomyces cerevisiae*, aiming to uncover the network's generic large-scale properties and the impact of the proteins' function and cellular localization on the network topology. They show how each database supports a scale-free, topology with hierarchical modularity, indicating that these features represent a robust and generic property of the PINs. They also find strong correlations between the network's structure and the functional role and sub-cellular localization of its protein constituents, concluding that most functional and/or localization classes appear as relatively segregated sub-networks of the full PIN. The uncovered systematic differences between the four protein interaction databases reflect their relative coverage for different functional and localization classes and provide a guide for their utility in various bioinformatics studies. Spirin *et al.* [118] discovered, by analyzing the structure of the yeast PIN, molecular modules that are densely connected within themselves but sparsely connected with the rest of the network. Comparison with experimental data and functional annotation of genes showed two types of modules: protein complexes (splicing machinery, transcription factors, etc.) and dynamic functional units (signaling cascades, cell-cycle regulation, *etc.*). Discovered modules are highly statistically significant, as is evident from comparison with random graphs, and are robust to noise in the data. The results of all these works are consistent with those observed previously by Jeong *et al.* [119]. Their study of yeast PIN shown that the probability that a given yeast protein interacts with k other yeast proteins follows a power law with an exponential cut-off at $k_c \approx 20$, a topology that is also shared by the PIN of the bacterium *Helicobacter pylori* (studied in [120]). This indicates that the PIN in these two separate organisms forms a highly inhomogeneous scale-free network in which a few highly connected proteins play a central role in mediating interactions among numerous, less connected proteins. An important known consequence of the inhomogeneous structure is the network's simultaneous tolerance to random errors, coupled with fragility against the removal of the most connected nodes. Thus, the most highly connected proteins in the cell would be the most important for its survival. In spite of all these results, the study of the topological and statistical properties of budding yeast and human PINs carried out by Hase *et al.* [121] revealed that they are scale-rich and configured as highly optimized tolerance networks that are similar to the router-level topology of the Internet. This is

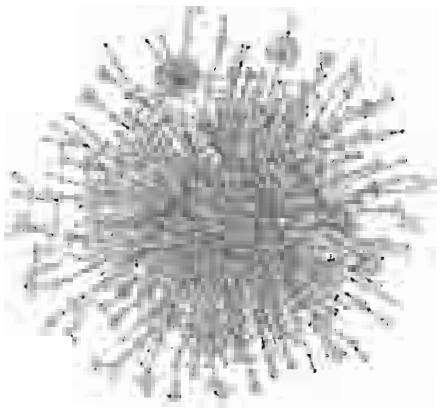


Fig. (8). Yeast protein interaction network.

different from claims that such networks are scale-free and configured through simple preferential-attachment processes. Further analysis revealed that there are extensive interconnections among middle-degree nodes (defined in this work as nodes with degrees within the range of 6-38 in yeast and of 6-30 in humans) that form the backbone of the networks. Degree distributions of essential genes, synthetic lethal genes, synthetic sick genes, and human drug-target genes indicate that there are advantageous drug targets among nodes with middle- to low-degree nodes (nodes with degrees of less than 5). Such network properties provide the rationale for combinatorial drugs that target less prominent nodes to increase synergetic efficacy and create fewer side effects.

In order to elucidate cellular machinery on a global scale, Sharan *et al.* [122] performed a multiple comparison of the PINs of *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Saccharomyces cerevisiae*. This comparison integrated protein interaction and sequence information to reveal 71 network regions that were conserved across all three species and many exclusive to the metazoans. They used this conservation, and found statistically significant support for 4,645 previously undescribed protein functions and 2,609 previously undescribed protein interactions. They tested 60 interaction predictions for yeast by two-hybrid analysis, confirming approximately half of these. Significantly, many of the predicted functions and interactions would not have been identified from sequence similarity alone, demonstrating that network comparisons provide essential biological information beyond what is gleaned from the genome.

5.4. Yeast Interaction Networks

The interaction networks of model eukaryotes such as yeast have been analyzed extensively. Because of that, in this work we have included a specific section devoted to yeast networks. DNA replication-division cycle in all eukaryotic cells is controlled by a common set of proteins interacting with each other to form a large and complex protein-protein interaction network (PIN), by a common set of rules, said Nurse [123]. However, each particular organism is characterized by its own interactions of proteins, leading to its own particularities of cell growth and division. In Fig. (**8**) we show the PIN of budding yeast. Generally, we have the same common genes and proteins and general dynamical principles that lead to the replication and division of the genome from mother cell to daughter, but we do not know for sure which parts of the common machinery are functioning in any given cell type, given the genetic background and developmental stage of an organism [124].

In fact, very recently Nandy *et al.* [125] published a work entitled: Reconstruction of the yeast protein-protein interaction network involved in nutrient sensing and global metabolic regulation. The authors noted that several PIN studies have been performed for the yeast *Saccharomyces cerevisiae* (budding yeast) using different high-throughput experimental techniques. All these results are collected in the BioGRID database and the SGD database provide detailed annotation of the different proteins. Despite the value of BioGRID for studying protein-protein interactions (PPIs), there is a need for manual curation of these interactions in order to remove false positives. In their work, these authors describe an annotated reconstruction of the PIN around four key nutrient-sensing and metabolic regulatory signal transduction pathways (STP) operating in yeast. The reconstructed STP network includes a full PIN including the key nodes Snf1, Tor1, Hog1 and Pka1. The network includes a total of 623 structural open reading frames (ORFs) and 779 PPIs. A number of proteins were identified having PPIs with more than one of the protein kinases. The fully reconstructed interaction network includes all the information available in separate databases for all the proteins included in the network (nodes) and for all the interactions between them (edges). The annotated information is readily available utilizing the functionalities of network modeling tools such as Cytoscape and CellDesigner. The reported fully annotated interaction model serves as a platform for integrated systems biology studies of nutrient sensing and regulation in yeast. Furthermore, they proposed this annotated reconstruction as a first step towards generation of an extensive annotated PIN of signal transduction and metabolic regulation in yeast.

In addition, Breitkreutz *et al.* [126] studied a global protein kinase and phosphatase interaction network in yeast. The interactions of protein kinases and phosphatases with their regulatory subunits and substrates underpin cellular regulation. These authors identified a kinase and phosphatase interaction (KPI) network of 1844 interactions in budding yeast by mass spectrometric analysis of protein complexes. The KPI network contained many dense local regions of interactions that suggested new functions. Notably, the cell cycle phosphatase Cdc14 associated with multiple kinases that revealed roles for Cdc14 in mitogen-activated protein kinase signaling, the DNA damage response, and metabolism, whereas interactions of the target of rapamycin complex 1 (TORC1) uncovered new effector kinases in nitrogen and carbon metabolism. An extensive backbone of kinase-kinase interactions cross-connects the proteome and may serve to coordinate diverse cellular responses.

Before, in 2008, Shi *et al.* [127] published a PIN analysis for yeast integral membrane protein. Although the yeast is the best exemplified single-celled eukaryote, the vast numbers of PPIs of integral membrane proteins of yeast have not been characterized by experiments. Here, based on the kernel method of Greedy Kernel Principal Component analysis plus Linear Discriminant Analysis, they identified 300 protein-protein interactions involving 189 membrane proteins and get the outcome of a highly connected PIN. Furthermore,

they study the global topological features of integral membrane PIN of yeast. These results give the comprehensive description of protein-protein interactions of integral membrane proteins and reveal global topological and robustness of the interactome network at a system level. This work represents an important step towards a comprehensive understanding of yeast protein interactions.

Lastly, we would like to cite here the work after Lin *et al.* [128] with title: *A comprehensive synthetic genetic interaction network governing yeast histone acetylation and deacetylation.* Histone acetylation and deacetylation are among the principal mechanisms by which chromatin is regulated during transcription, DNA silencing, and DNA repair. The authors analyzed patterns of genetic interactions uncovered during comprehensive genome-wide analyses in yeast to probe how histone acetyltransferase (HAT) and histone deacetylase (HDAC) protein complexes interact. The genetic interaction data unveil an underappreciated role of HDACs in maintaining cellular viability, and led to show that deacetylation of the histone variant Htz1p at Lys 14 is mediated by Hda1p. Studies of the essential nucleosome acetyltransferase of H4 (NuA4) revealed acetylation-dependent protein stabilization of Yng2p, a potential nonhistone substrate of NuA4 and Rpd3C, and led to a new functional organization model for this critical complex. They also found that DNA double-stranded breaks (DSBs) result in local recruitment of the NuA4 complex, followed by an elaborate NuA4 remodeling process concomitant with Rpd3p recruitment and histone deacetylation. These new characterizations of the HDA and NuA4 complexes demonstrate how systematic analyses of genetic interactions may help illuminate the mechanisms of intricate cellular processes.

5.5. Metabolic Pathway Networks

Study of the metabolic pathways from the point of view of the network theory allow us to understand the mechanisms implied in the normal functioning of the organisms and to identify reactions that are crucial for their survival, being this an important step in the strategies followed to design new drugs. Jeong *et al.* [129] presented a systematic comparative mathematical analysis of the metabolic networks of 43 organisms representing all three domains of life. Despite significant variation in their individual constituents and pathways, these metabolic networks have the same topological scaling properties and show striking similarities to the inherent organization of complex non-biological systems. This may indicate that metabolic organization is not only identical for all living organisms, but also complies with the design principles of robust and error-tolerant scale-free networks, and may represent a common blueprint for the large-scale organization of interactions among all cellular constituents. Guimera & Nunes-Amaral [130] demonstrated that it is possible to find functional modules in complex networks, and classify nodes into universal roles according to their pattern of intra and inter-module connections. They used their method to analyze the metabolic networks of 12 organisms from three different super-kingdoms and found that, typically, 80 % of the nodes are only connected to other nodes within their respective modules, and that nodes with different roles are affected by different evolutionary constraints and pressures. Remarkably, they find that metabolites that participate in only a few reactions but connect with different modules are more conserved than hubs whose links are mostly within a single module. Wagner & Fell [131] have undertaken a graph theoretical analysis of the *Escherichia coli* metabolic network and find that this network is a small-world graph. Moreover, the connectivity of the metabolites follows a power law. This provides an objective criterion for the centrality of the tricarboxylic acid cycle to metabolism. The small-world architecture may serve to minimize transition times between metabolic states, and contains evidence about the evolutionary history of metabolism. Duarte *et al.* [132] have manually reconstructed the global human metabolic network based on build 35 of the genome annotation and a comprehensive evaluation of > 50 years of legacy data (*i.e.*, bibliomic data). They describe the reconstruction process and demonstrate how the resulting genome-scale (or global) network can be used (*i*) for the discovery of missing information, (*ii*) for the formulation of an *in silico* model, and (*iii*) as a structured context for analyzing high-throughput biological data sets. The evaluation of the literature revealed many gaps in the current understanding of human metabolism that require future experimental investigation. Mathematical analysis of network structure elucidated the implications of intracellular compartmentalization and the potential use of correlated reaction sets for alternative drug target identification. Integrated analysis of high-throughput data sets within the context of the reconstruction enabled a global assessment of functional metabolic states. These results highlight some of the applications enabled by the reconstructed human metabolic network. Patil & Nielsen [133] developed an algorithm (Fig. **9**) that is based on hypothesis-driven data analysis to uncover the transcriptional regulatory architecture of metabolic networks. By using information on the metabolic network topology from genome-scale metabolic reconstruction, they show that it is possible to reveal patterns in the metabolic network that follow a common transcriptional response. Thus, the algorithm enables identification of so-called reporter metabolites (metabolites around which the most significant transcriptional changes occur) and a set of connected genes with significant and coordinated response to genetic or environmental perturbations. They found that cells respond to perturbations by changing the expression pattern of several genes involved in the specific part(s) of the metabolism in which a perturbation is introduced. These changes then are propagated through the metabolic network because of the highly connected nature of metabolism.

Rahman & Schomburg [134] developed a method ('load points') to rank the enzymes/metabolites in the metabolic network and proposed a model to determine and rank the biochemical lethality in metabolic networks (enzymes/metabolites) through 'choke points'. Based on an extended form of the graph theory model of metabolic networks, metabolite structural information was used to calculate the k-shortest paths between metabolites (the presence of more than one competing path between substrate and product). On the basis of these paths and connectivity information, load points were calculated and used to empirically rank the importance of metabolites/enzymes in the metabolic network. The load point analysis emphasizes the role that the biochemical structure of a metabolite, rather than its connec

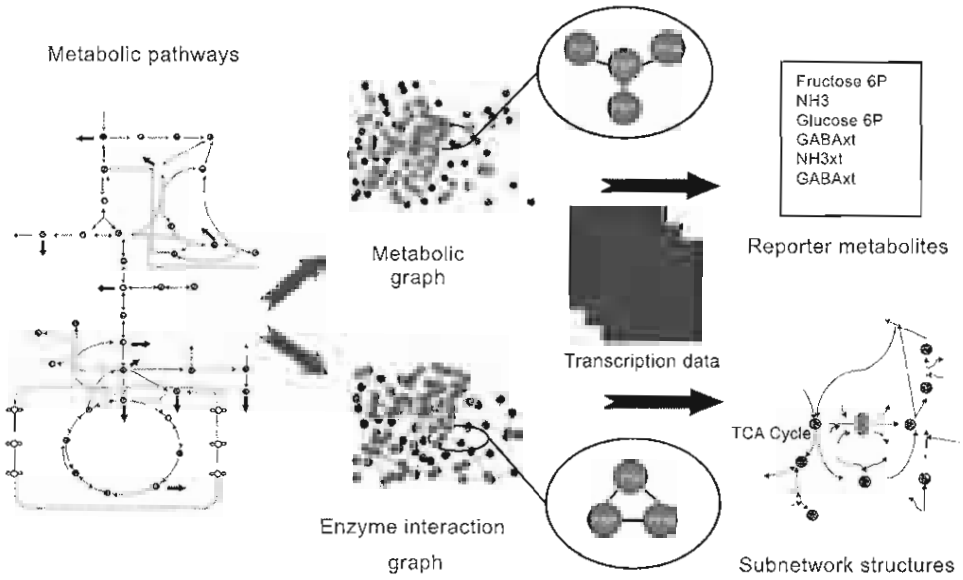


Fig. (9). Illustration of the proposed algorithm for identifying reporter metabolites and subnetwork structures signifying transcriptionally regulated modules. A metabolic network (set of reactions) is converted to bipartite (metabolic) and unipartite (enzyme-interaction) graph representations. Gene-expression data from a particular experiment then is used to identify highly regulated metabolites (reporter metabolites) and significantly correlated subnetworks in the enzyme-interaction graph. Reproduced from [133].

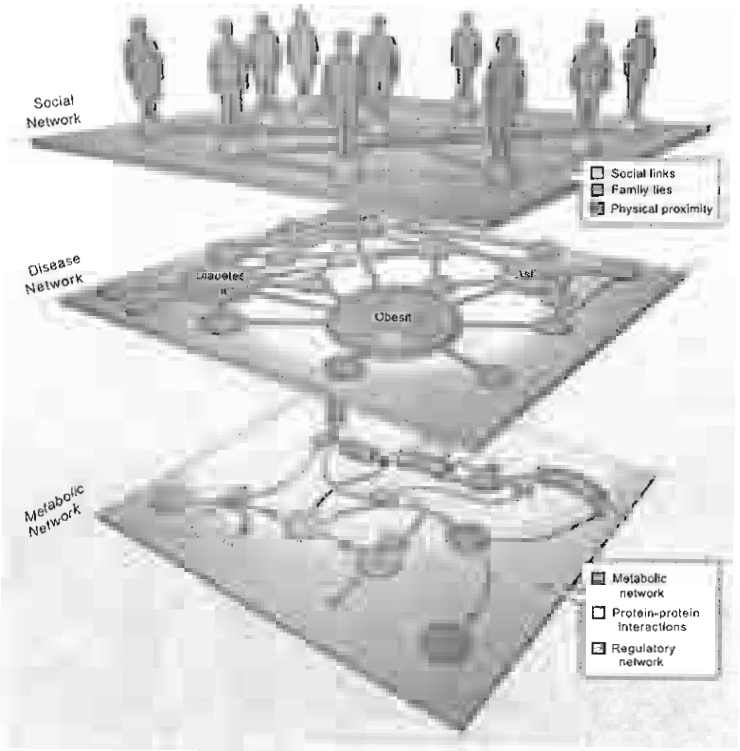


Fig. (10). Complex networks of direct relevance to study of diseases. Although they are often treated separately, most human diseases are not independent of each other. Many diseases are associated with the breakdown of functional modules that are best described as sub-networks of a complex network connecting many cellular components. Therefore, an understanding of the functionally relevant genetic, regulatory, metabolic, and protein–protein interactions in a cellular network will play an important role in understanding the pathophysiology of human diseases (bottom layer). One way to visualize the ensuing potential interrelationship among human diseases is to construct a disease network (middle layer) in which two diseases are connected each other if they have a common genetic or functional origin. For example, on the basis of our current knowledge of disease genes, obesity is connected to at least seven other diseases such as diabetes, asthma, and insulin resistance, since genes associated with these diseases are known to affect obesity as well. The third network of key importance to human disease is the social network, which encompasses all human-to-human interactions (e.g., familial, friendship, sexual, and proximity-based contacts) that play a role in the spread of pathogens (top layer). Efforts to understand the interactions between the cellular, disease, and social networks are part of network medicine, which aims to quantify the complex interlinked factors that may contribute to individual diseases. Reproduced from [136].

tivity (hubs), plays in the conversion pathway. In order to identify potential drug targets (based on the biochemical lethality of metabolic networks), the concept of choke points and load points was used to find enzymes (edges) which uniquely consume or produce a particular metabolite (nodes). A non-pathogenic bacterial strain *Bacillus subtilis* 168 (lactic acid producing bacteria) and a related pathogenic bacteria strain *Bacillus anthracis Sterne* (avirulent but toxigenic strain, producing the toxin Anthrax) were selected as model organisms. The choke point strategy was implemented on the pathogen bacterial network of *B. anthracis Sterne*. Potential drug targets were proposed based on the analysis of the top 10 choke points in the bacterial network. A comparative study between the reported top 10 bacterial choke points and the human metabolic network was performed. Further biological inferences were made on results obtained by performing a homology search against the human genome. Lemke *et al.* [135] studied the annotated genome sequence of the microbe *Escherichia coli* and propose a general quantitative definition of enzyme importance in a metabolic network. Using a graph analysis of its metabolism, they relate the extent of the topological damage generated in the metabolic network by the deletion of an enzyme to the experimentally determined viability of the organism in the absence of that enzyme. They show that the network is robust and that the extent of the damage relates to enzyme importance. They predict that a large fraction (91 %) of enzymes causes little damage when removed, while a small group (9 %) can cause serious damage. Experimental results confirm that this group contains the majority of essential enzymes. The results may reveal a universal property of metabolic networks.

6. DISEASE NETWORKS

In the previous section we have seen how the molecular networks can be used in drug design and in understanding the functional and organizational principles of molecular and metabolic processes. In this section we are going to speak about the diseases from the point of view of the network analysis. Barabási [136] proposed that there are three types of complex networks of direct relevance to study of diseases (Fig. **10**): Metabolic networks, disease networks and social networks. The metabolic networks describe the functionally relevant genetic, regulatory, metabolic, and protein–protein interactions; the disease networks describe the relationships among diseases and the social networks describe familial, friendship, sexual and proximity-based contacts.

Loscalzo *et al.* [137] answered to the question: What is the benefit of a network analysis of disease and its treatment? First, systems-based network analysis can identify those determinants (nodes) or combinations of determinants that strongly influence network behavior and disease expression or phenotype. Second, these regulatory determinants may not always be obvious from reductionist principles, and, thus, the analysis provides unique insight into disease mechanism and potential therapeutic targets. Third, network analysis of disease gives one the opportunity to consider with quantitative rigor the relationships within the network genome, environmental exposures, and environmental effects on the proteome (posttranslational proteome) that define the specific pathophenotype. In this construct, disease can be considered the result of a modular collection of genomic, proteomic, metabolomic, and environmental networks that interact to yield the pathophenotype. Fourth, disease network analysis ultimately provides a mechanistic basis for defining phenotypic differences among individuals with the same disease through consideration of unique genetic and environmental factors that govern intermediate phenotypes contributing to disease expression. Lastly, disease network analysis offers a unique method for identifying therapeutic targets or combinations of targets that can alter disease expression.

6.1. Disease-Disease Networks

There are no clear boundaries between many diseases, as diseases can have multiple causes and can be related through several dimensions. From a genetic perspective, a pair of diseases can be related because they have both been associated with the same gene, whereas from a proteomic perspective diseases can be related because disease associated proteins act on the same pathway. Hidalgo *et al.* [138] introduced a phenotypic database summarizing correlations obtained from the disease history of more than 30 million patients in a Phenotypic Disease Network (PDN). They presented evidence that the structure of the PDN is relevant to the understanding of illness progression by showing that (*i*) patients develop diseases close in the network to those they already have; (*ii*) the progression of disease along the links of the network is different for patients of different genders and ethnicities; (*iii*) patients diagnosed with diseases which are more highly connected in the PDN tend to die sooner than those affected by less connected diseases; and (*iv*) diseases that tend to be preceded by others in the PDN tend to be more connected than diseases that precede other illnesses, and are associated with higher degrees of mortality. Their findings show that disease progression can be represented and studied using network methods, offering the potential to enhance our understanding of the origin and evolution of human diseases.

6.2. Diseasome Networks

A diseasome can be defined as a network of disorders and disease genes linked by known disorder-gene associations (Fig. **11**). Goh *et al.* [139] found that essential human genes are likely to encode hub proteins and are expressed widely in most tissues. This suggests that disease genes also would play a central role in the human interactome. In contrast, they found that the vast majority of disease genes are nonessential and show no tendency to encode hub proteins, and their expression pattern indicates that they are localized in the functional periphery of the network. A selection-based model explains the observed difference between essential and disease genes and also suggests that diseases caused by somatic mutations should not be peripheral, a prediction they confirm for cancer genes.

6.3. Protein-Disease Networks

During a decade of proof-of-principle analysis in model organisms, protein networks have been used to further the study of molecular evolution, to gain insight into the robustness of cells to perturbation, and for assignment of new protein functions. Following these analyses, and with the recent

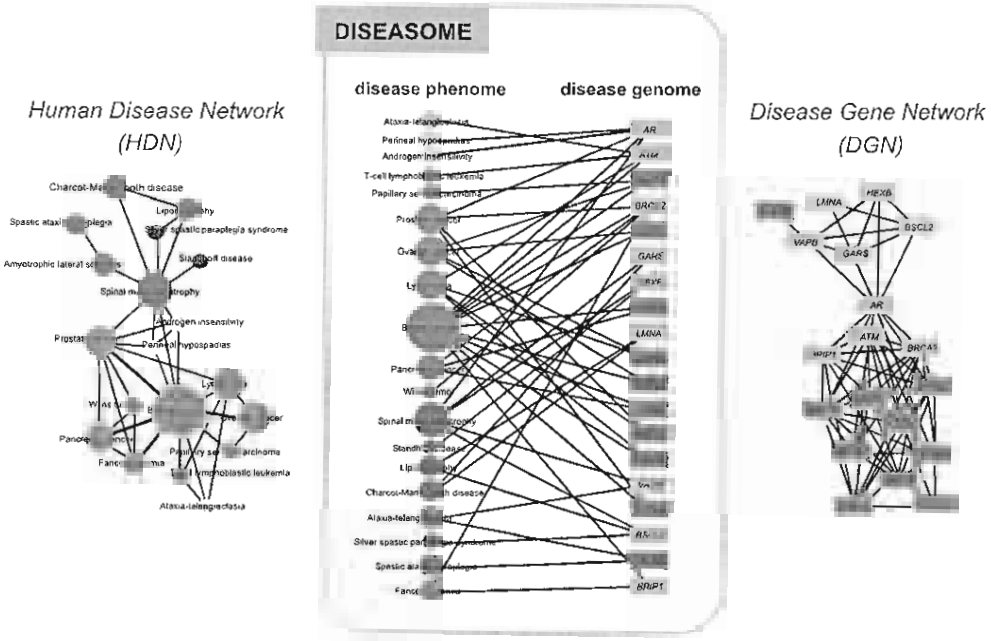


Fig. (11). Construction of the diseasome bipartite network (*Center*). A small subset of OMIM-based disorder–disease gene associations, where circles and rectangles correspond to disorders and disease genes, respectively. A link is placed between a disorder and a disease gene if mutations in that gene lead to the specific disorder. The size of a circle is proportional to the number of genes participating in the corresponding disorder, and the color corresponds to the disorder class to which the disease belongs. (*Left*) The HDN projection of the diseasome bipartite graph, in which two disorders are connected if there is a gene that is implicated in both. The width of a link is proportional to the number of genes that are implicated in both diseases. For example, three genes are implicated in both breast cancer and prostate cancer, resulting in a link of weight three between them. (*Right*) The DGN projection where two genes are connected if they are involved in the same disorder. The width of a link is proportional to the number of diseases with which the two genes are commonly associated. Reproduced from [139].

rise of protein interaction measurements in mammals, protein networks are increasingly serving as tools to unravel the molecular basis of disease [140]. In order to discover, in high throughput, the molecular underpinnings of poorly characterized diseases, Sam *et al.* [141] presented a statistical method to identify shared protein interaction network(s) between diseases. Integrating (*i*) a protein interaction network with (*ii*) disease to protein relationships derived from mining Gene Ontology annotations and the biomedical literature with natural language understanding (PhenoGO), they identified protein-protein interactions that were associated with pairs of diseases and calculated the statistical significance of the occurrence of interactions in the protein interaction knowledgebase. Significant correlations between diseases and shared protein networks were identified and evaluated, demonstrating the high precision of the approach and correct non-trivial predictions, signifying the potential for discovery. In conclusion, they demonstrate that the associations between diseases are directly correlated to their underlying protein-protein interaction networks, possibly providing insight into the underlying molecular mechanisms of phenotypes and biological processes disrupted in related diseases. Liu *et al.* [142] studied the type 2 diabetes mellitus using a network-based analysis methodology, identifying two sets of genes associated with insulin signaling and a network of nuclear receptors, which are recurrent in a statistically significant number of diabetes and insulin resistance models and transcriptionally altered across diverse tissue types. They additionally identified a PIN between members from the two gene sets that may facilitate signaling between them. The results illustrate the benefits of integrating high-throughput microarray studies, together with PINs, in elucidating the underlying biological processes associated with a complex disorder. Lee *et al.* [143] constructed a bipartite human disease association network in which nodes are diseases and two diseases are linked if mutated enzymes associated with them catalyze adjacent metabolic reactions. They find that connected disease pairs display higher correlated reaction flux rate, corresponding enzyme-encoding gene co-expression, and higher co-morbidity than those that have no metabolic link between them. Furthermore, the more connected a disease is to other disease; the higher is its prevalence and associated mortality rate. The network topology-based approach also helps to uncover potential mechanisms that contribute to their shared pathophysiology. Thus, the structure and modeled function of the human metabolic network can provide insights into disease comorbidity, with potentially important consequences for disease diagnosis and prevention. The analysis performed by Xu & Li [144] have revealed that the hereditary disease-genes ascertained from the Online Mendelian Inheritance in Man database (OMIM) in the literature-curated PINs are characterized by a larger degree, tendency to interact with other disease-genes, more common neighbors and quick communication to each other whereas those properties could not be detected from the network identified from high-throughput yeast two-hybrid mapping approach and predicted interactions (PDT) PINs. K-nearest neighbors classifier based on those features was created and on average gained overall prediction accuracy of 0.76 in cross-validation test. Then the classifier was applied

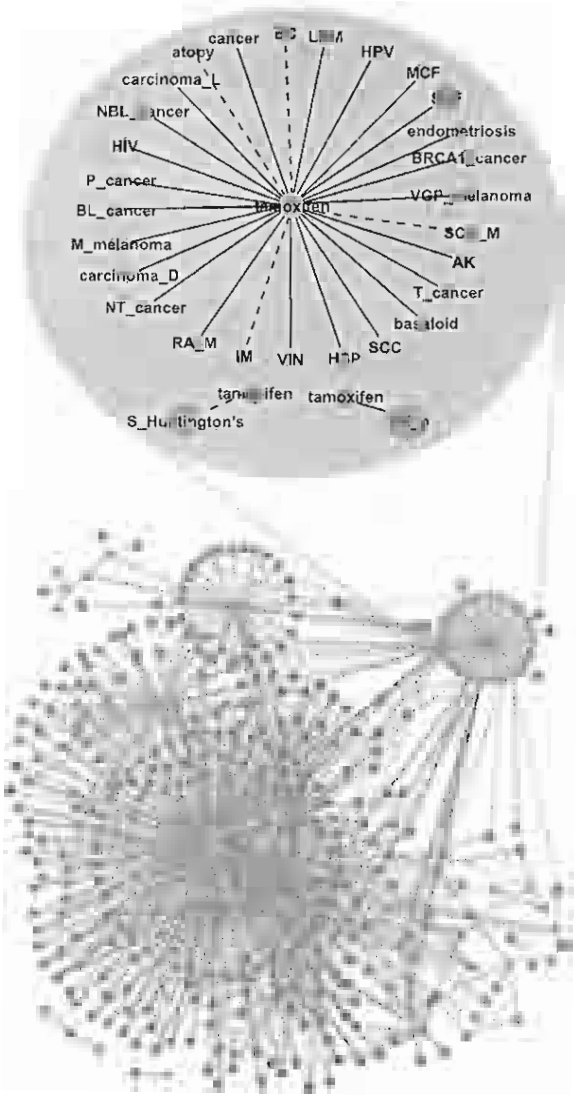


Fig. (12). Disease-drug network. This disease-drug network contains a total of 49 diseases in dark cyan nodes, 213 drugs in gold, and 906 connections. The size of the nodes is proportional to the number of links. Positive matches are shown by solid lines and negative relationships by dotted lines. Multiple nodes with the same descriptive name exist because the corresponding profiles were generated under different conditions or studies. Reproduced from [147].

to 5,262 genes on human genome and predicted 178 novel disease-genes. Some of the predictions have been validated by biological experiments. Jonsson & Bates [145] shown that human proteins translated from known cancer genes exhibit a network topology that is different from that of proteins not documented as being mutated in cancer. In particular, cancer proteins show an increase in the number of proteins they interact with. They also appear to participate in central hubs rather than peripheral ones, mirroring their greater centrality and participation in networks that form the backbone of the proteome. Moreover, they show that cancer proteins contain a high ratio of highly promiscuous structural domains, i.e., domains with a high propensity for mediating protein interactions. These observations indicate an underlying evolutionary distinction between the two groups of proteins, reflecting the central roles of proteins, whose mutations lead to cancer. The aim of the study carried out by Goni *et al.* [146] was to assess whether the parameters of degree and betweenness are properties that differentiate between implicated (seed-proteins) and non-implicated nodes (neighbors) in Multiple Sclerosis (MS) and Alzheimer disease (AD). They used experimentally validated PPI information to obtain the neighbors for each seed group and they studied these parameters in four networks: MS-blood network; MS-brain network; AD-blood network; and AD-brain network. Specific features of seed-proteins were revealed, whereby they displayed a lower average degree in both diseases and tissues, and a higher betweenness in AD-brain and MS-blood networks. Additionally, the heterogeneity of the processes involved indicate that these findings are not pathway specific but rather that they are spread over different pathways. In summary, their findings show differential centrality properties of proteins whose gene expression is impaired in neurodegenerative diseases.

6.4. Drug-Disease Networks

The traditional view of drug action on disease as a 'key' fitting into the 'lock' is certainly over-simplified and has been challenged by a growing body of evidence showing that there are many keys for each lock and a single key can fit multiple locks [202]. Hu & Agarwal [147] performed a systematic, large-scale analysis of genomic expression profiles of human diseases and drugs to create a disease-drug network (Fig. **12**). A network of 170,027 significant interactions was extracted from the ~24.5 million comparisons between ~7,000 publicly available transcriptomic profiles. The network includes 645 disease-disease, 5,008 disease-drug, and 164,374 drug-drug relationships. At least 60% of the disease-disease pairs were in the same disease area as determined by the Medical Subject Headings disease classification tree. The remaining can drive a molecular level nosology by discovering relationships between seemingly unrelated diseases, such as a connection between bipolar disorder and hereditary spastic paraplegia, and a connection between actinic keratosis and cancer. Among the 5,008 disease-drug links, connections with negative scores suggest new indications for existing drugs, such as the use of some antimalaria drugs for Crohn's disease, and a variety of existing drugs for Huntington's disease; while the positive scoring connections can aid in drug side effect identification, such as tamoxifen's undesired carcinogenic property. From the ~37K drug-drug relationships, they discover relationships that aid in target and pathway deconvolution, such as 1) KCNMA1 as a potential molecular target of lobeline, and 2) both apoptotic DNA fragmentation and G2/M DNA damage checkpoint regulation as potential pathway targets of daunorubicin.

6.5. Disease Spreading Networks

Understanding the nature of animal contact patterns is crucial for predicting the impact of future pandemics and devising effective control measures. Individuals in a population susceptible to a disease may be represented as vertices in a network, with the edges that connect vertices representing social and/or spatial contact between individuals. Shirley & Rushton [148] created networks with six different patterns of connection between vertices. Both scale-free networks and random graphs showed a different response in path level

to increasing levels of clustering than regular lattices. Clustering promoted short path lengths in all network types, but randomly assembled networks displayed a logarithmic relationship between degree and path length; whereas this response was linear in regular lattices. In all cases, small-world models, generated by rewiring the connections of a regular lattice, displayed properties, which spanned the gap between random and regular networks. Simulation of a disease in these networks showed a strong response to connectivity pattern, even when the number of edges and vertices were approximately equal. Epidemic spread was fastest, and reached the largest size, in scale-free networks, then in random graphs. Regular lattices were the slowest to be infected, and rewired lattices were intermediate between these two extremes. Scale-free networks displayed the capacity to produce an epidemic even at a likelihood of infection, which was too low to produce an epidemic for the other network types.

Many epidemiological models make simplifying assumptions about the patterns of disease-causing interactions among hosts. In particular, homogeneous-mixing models assume that all hosts have identical rates of disease-causing contacts. In recent years, several network-based approaches have been developed to explicitly model heterogeneity in host contact patterns. Bansal *et al.* [149] used a network perspective to quantify the extent to which real populations depart from the homogeneous-mixing assumption, in terms of both the underlying network structure and the resulting epidemiological dynamics. They found that human contact patterns are indeed more heterogeneous than assumed by homogeneous-mixing models, but are not as variable as some have speculated. They then evaluated a variety of methodologies for incorporating contact heterogeneity, including network-based models and several modifications to the simple SIR compartmental model. They conclude that the homogeneous-mixing compartmental model is appropriate when host populations are nearly homogeneous, and can be modified effectively for a few classes of non-homogeneous networks. In general, however, network models are more intuitive and accurate for predicting disease spread through heterogeneous host populations. Read *et al.* [150] presented the results from a detailed diary-based survey of casual (conversational) and close contact (physical) encounters made by a small peer group of 49 adults who recorded 8,661 encounters with 3,528 different individuals over 14 non-consecutive days. They found that the stability of interactions depends on the intimacy of contact and social context. Casual contact encounters mostly occur in the workplace and are predominantly irregular, while close contact encounters mostly occur at home or in social situations and tend to be more stable. Simulated epidemics of casual contact transmission involve a large number of non-repeated encounters, and the social network is well captured by a random mixing model. However, the stability of the social network should be taken into account for close contact infections. Shirley & Rushton [151] used a graph-theoretical approach to investigate the properties of the observed network of disease transmission in the 2001 foot-and-mouth epidemic in the United Kingdom. This analysis revealed both global and local heterogeneity in the contact pattern between the infected premises in the first 3 weeks of the disease. In particular, the global heterogeneity contributed to the failure of the culling strategy imposed by the UK government. However, a more effective strategy targeting selective deletion of key premises in the network was not available once the epidemic had begun. They recommend that post-hoc analyses of this sort should become part of preventative and proactive policy rather than part of a reaction to an ongoing crisis. Colizza *et al.* [152] present a stochastic computational framework for the forecast of global epidemics that considers the complete worldwide air travel infrastructure complemented with census population data. The authors address two basic issues in global epidemic modeling: (*i*) they study the role of the large scale properties of the airline transportation network in determining the global diffusion pattern of emerging diseases; and (*ii*) they evaluate the reliability of forecasts and outbreak scenarios with respect to the intrinsic stochasticity of disease transmission and traffic flows. To address these issues they define a set of quantitative measures able to characterize the level of heterogeneity and predictability of the epidemic pattern. These measures may be used for the analysis of containment policies and epidemic risk assessment.

Fasciolosis is a parasitic infection caused by *Fasciola hepatica* that has become an important cause of lost productivity in livestock worldwide. Effective control of fasciolosis is difficult, especially in milking cows, which can only be treated during the dry period. González-Díaz *et al.* [153] constructed a network for fasciolosis spreading in Galicia. They also calculated many centrality measures for all the nodes of the new network (livestock farms). Lastly, using these measures of landscape network structure as inputs, they seek a Quantitative Structure-Property Relationship (QSPR) model. This QSPR model may predict the prevalence of disease, in the former or new farms, after or in absence of different medical treatments based only on details retrieved from GIS (location and altitude). The study may have predictive value for the positioning of new farms with lower risk of infection or in managing cattle during infections.

7. REVIEW OF CLASSIC TIS FOR COMPLEX NETWORKS.

7.1. Classic TIs for Full Graphs

The distribution of systems that are amenable to representation and study with graph or network theory in nature is so vast that many authors consider the network theory a science [154]. In any case, describing the connectivity or topology of chemical and/or biological systems with discrete structures using graphs or networks is a gateway to the introduction of mathematical and bioinformatics tools in Proteomics [155]. Graphs and networks are simple objects that contain at least nodes and edges. Each node represents one part of a complex system and the edges represent geometric and/or functional relationships between these parts. The use of graphical approaches to study complicated biological systems can provide an intuitive picture and help to gain useful insights into such systems. In Proteomics, amino acids, proteins, electrophoresis spots, polypeptidic fragments, or more complex objects can play the role of nodes. In these cases, the system may be a protein, Protein Interaction Networks (PINs), a 2D proteome electrophoresis map, or a serum

plasma proteome mass spectrum for a patient, respectively. All these graphs or networks can be numerically described using the so-called Topological Indices (TIs) [65, 156]. The transformation of graphs (a picture) into TIs (numbers) allows not only the storage, manipulation, comparison and retrieval of information in Proteomics, but also enables the search for quantitative relationships between system structure and function. The search for structure-activity or quantitative-structure-activity relationships (SAR/QSAR) for small-sized drugs is a very active field in biomedical sciences [157]. When the drug is represented as a graph it is possible to use graph TIs of molecules to relate the chemical structure with the activity. However, one can consider numerous variations to this classic strategy beyond the world of small molecules and covering not only macromolecules, such as DNA and proteins, but brain cortex, population sociology and other complex systems [158-164]. In any case, all these systems follow to a large extent a series of common stages, irrespective of the class of objects under investigation (molecular graphs or complex networks).

Inspection of a classic QSAR workflow clearly shows that the collection of biological activity data, system structure codification and subsequent data analysis are three of the most important bottlenecks. In particular, the encryption of the system structure information with different graph or network TIs has been a very active field of research. Several TIs have been introduced up to date. A compilation by Todeschini and Consonni systematizes more than 1,600 molecular descriptors for small-molecule drug discovery, including several TIs. Some of these are redundant in some way or have topics in common [65, 156]. For instance, many researchers define TIs for graphs or networks using vector-matrix-vector procedures, a fact that indicates significant similarities between these systems [165-167]. Indeed, the first TIs ever defined in a chemical context, the Wiener index W (se next equations), are in a vector-matrix-vector form. In addition, several other TIs can be written in a vector-matrix-vector form and these include [168]:

1. Zagreb indices M1 and M2
2. Harary number H
3. Randič connectivity index χ
4. Valence connectivity index χ^v
5. Marrero–Ponce quadratic or linear forms
6. Balaban index J
7. Broto–Moreau autocorrelation ATS_d
8. Graph TIs for Markov matrices reported by González-Díaz *et al.*

$$W = \frac{1}{2}\left(\mathbf{u}\cdot\mathbf{D}\cdot\mathbf{u}^T\right) \quad M_1 = \mathbf{v}\cdot\mathbf{A}\cdot\mathbf{u}^T \quad M_2 = \frac{1}{2}\left(\mathbf{v}\cdot\mathbf{A}\cdot\mathbf{v}^T\right)$$

$$H = \frac{1}{2}\left(\mathbf{u}\cdot\mathbf{D}^{k}\cdot\mathbf{u}^T\right) \qquad \chi = \mathbf{v}^{i}\cdot\mathbf{A}\cdot\mathbf{v}^{iT} \quad \chi^v = \mathbf{v}^{ii}\cdot\mathbf{A}\cdot\mathbf{v}^{iiT}$$

$$q_k(X) = \mathbf{w}\cdot\mathbf{M}\cdot\mathbf{w}^T \qquad f_k(X) = \mathbf{w}\cdot\mathbf{M}\cdot\mathbf{u}^T \quad s_k(X) = \mathbf{w}\cdot\mathbf{S}_k\cdot\mathbf{w}^T$$

$$J = \frac{1}{2}\cdot C\cdot\left(\mathbf{d}^{i}\cdot\mathbf{A}\cdot\mathbf{d}^{iT}\right) \quad ATS_d = \mathbf{w}\cdot{}^{m}\mathbf{B}\cdot\mathbf{w}^T \quad \xi_k = \mathbf{p}\cdot{}^{1}\Pi\cdot\mathbf{w}^T$$

All the vectors and matrices used in previous expressions have been exhaustively explained in the literature and the reader can find further details there [165, 166, 169, 170]. In the case of Marrero–Ponce TIs, there is a more elaborate approach in terms of algebraic space, while invariants reported by González-Díaz *et al.* are explained through Markov Chains theory. In any case, this common feature leads one to expect that a large number of researchers will no longer proceed with the development of new TIs. We share the same opinion in terms of the development of TIs for small-sized molecules. However, there are still some very important topics that have yet to be comprehensively covered by classical TIs or the most elaborated topographic indices (TPGIs). Consequently, renewed interest has recently been shown in the development and application of new TIs and TPGIs with the intention of extending the potential applications of QSAR approaches [65, 156, 167, 171, 172].

7.2. Centrality Measures for Nodes

We can use TIs to define centralities of nodes in many different biological technological, or social network-like systems. Centralities $C_t(j)$ of type t for a j^{th} node [173, 174] are numbers that numerically characterize a specific node, opposite to the TIs that describe properties of the entire graph/complex network. Network centralities are used to rank elements of a network according to a given importance concept. Generally, for each TI it is possible to calculate the corresponding centrality as a difference between the TI of the entire graph (G) and the TI for the graph G without a specific node j ($G - j$):

$$C_t(j) = TI(G) - TI(G - j)$$

The definition of new Centralities is an active field of research and new centralities such as sub-graph centrality have been introduced by Estrada [175].

8. MARKOV TIs FOR COMPLEX NETWORKS

8.1. Review of Recent Markov-Chain Based TIs

The classical TIs can include additional information such as Markov node linkage probability (p_{ij}) for any i, j nodes of a graph. Therefore, we have introduced several types of Markov TIs such as Markov-Shannon Entropy [176], Markov-Randic indices [177], or Markov-Harary numbers [47]. In these works we have used the Markov-TIs in order to compare several types of complex networks from different fields as Biology Linguistics, Technology, Sociology and Law by using the centralities calculated with a new tool, MCeConet. The results obtained have shown the usefulness of the Markov-TIs in network-based studies.

8.2. Definition of New Stochastic Spectral Moments TIs for Complex Networks

In the present work, we have constructed the classical Markov matrix ($^1\mathbf{\Pi}$) for each network as follows:

a) The link connectivities between the nodes of the networks generate the connectivity matrix, **C**, a $n \times n$ matrix with c_{ij} values of 1 (for connections between nodes i and j) or 0 (for non-connection);

b) The Markov matrix $\mathbf{\Pi}$ is built and contains the probability of the vertices (p_{ij}) based on **C**; p_{ij} are calculated by dividing

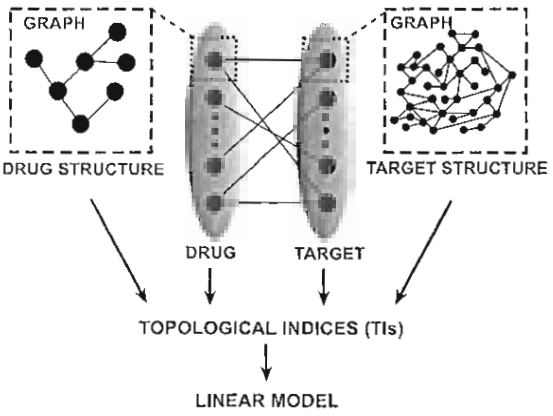


Fig. (13). Scheme of the development of a model that takes into account the drug and target structure and the information about the drug/target nodes.

each c_{ij} to the sum of the elements corresponding to i line in matrix **C**;

c) The probability matrix is raised to the power k (0-5), resulting $({}^{1}\Pi)^{k}$, which contains the elements ${}^{k}p_{ij}$;

d) The node linkage Markov probability (${}^{k}p_{ij}$) are used to calculate the Markov or stochastic spectral moments (π_k) for the entire graph and power k.

e) If we carry out the sum over a group or sub-set of nodes G we can calculate: a total TI if G include the n nodes, a local TI if G include some nodes, or an node centrality when G includes only the jth node.

f) MI-NODES (March-Inside node descriptors) is a GUI Python/wxPython application used for the calculation of a new class of Markov TIs or centralities of a network based on the Markov normalized node probabilities. It is an upgrade of the software MARCH-INSIDE.

$$\pi_k(G) = Tr\left[\left({}^{1}\Pi\right)^{k}\right]_{i=j\in G} = \sum_{i=j\in G}^{n} {}^{k}p_{ij}$$

9. STOCHASTIC SPECTRAL MOMENTS MODELS FOR COMPLEX NETWORKS

In this section we have evaluated the performance of the stochastic spectral moments in describing interactions between nodes of large networks from different fields. To do this, we have calculated the spectral moments by using Mi-Nodes software and then we have used the linear discriminant analysis (LDA) to find a simple model that reconstructs properly the links between nodes. This analysis was carried out by means of the STATISTICA software [178]. The graphical representation of the networks and calculation of the network descriptors were carried out by means of the Pajek and CentiBin software.

9.1. New mt-QSAR Model for Drug-Target Complex Network

As was seen in previous sections, the study of drug-target interaction networks is an important topic in drug development. Therefore it is interesting to find mathematical models that describe in a simple way the drug-target interactions. Here, we have developed a model that takes into account the drug and target structure and the information about the drug/target nodes in the studied network (Fig. **13**). This network was constructed using the data obtained from the Federal Food and Drug Administration (FDA) and the best model found was:

$$\begin{aligned} S(\text{drug-target}) &= 1.3887 \cdot \pi_2(\text{D.node}) - 0.5575 \cdot {}^{3}\pi_3(\text{D.C}_{\text{sat}}) - 0.4171 \cdot \pi_1(\text{D.C}_{\text{unsat}}) \\ &+ 0.8896 \cdot \pi_0(\text{D.Hal}) + 0.8010 \cdot \pi_0(\text{D.Het}) + 0.0195 \cdot \pi_1(\text{T.Total}) + 1.1342 \\ n &= 2,234 \quad \chi^2 = 1,204.779 \quad p < 0.001 \end{aligned}$$

Where, π_k is the stochastic spectral moment after k steps. The π_k values used to seek the equation have been calculated for sub-sets of nodes in different graphs. Some π_k values are for nodes of drugs in the drug-target network (D.node). Other π_k values are for nodes of atoms in the drug molecular graph: for saturated carbon atoms ($D.C_{sat}$), for unsaturated carbon atoms ($D.C_{unsat}$), for halogen atoms (D. Hal), and for heteroatoms (D.Het). We also included π_k values of nodes of all aminoacids in the protein structure network for drug target (T. Total). The parameter n is the number of cases used in the analysis. The values of the studied data were standardized previously to the LDA. The model correctly classified in train/cross validation 76.01/76.94% of the negative cases and 81,83/79.17% of the positive cases (interactions) with an accuracy of 77.89/77.68% (Table **5**). The data obtained from the model were used to reconstruct the observed network and to compare, by means of network descriptors, the observed and reconstructed networks (Table **6**).

9.2. New QSPR Model for Parasite-Host Networks

Due to the importance for the human health and the economy, much attention has been focused on the parasite-host interactions. The study of these interactions is essential to understand the role of phylogenetic and ecological factors on the parasite-host specificity [179-181] and to know how parasites affect the ecosystem functioning [182-184]. In order to construct the studied networks we used two sources of information. The first of them is the Interaction Web Database (IWDB) (http://www.nceas.ucsb.edu/interactionweb/index.html) that contains datasets on species interactions from several communities in different parts of the world. In particular we used the host-parasite dataset, composed of data belonging to studies about parasites (nematodes, acanthocephalans, cestodes, trematodes, monogeneans, leeches, copepods and branchiurans) and their hosts (fish) from 7 Canadian freshwater systems [185-190]. The second source of information is the Global Mammal Parasite Database (GMPD) (http://www.mammalparasites.org/), a compilation of records of parasites (helminths, protozoa, viruses, bacteria, arthropods and fungi) and their hosts (wild mammals) that have been documented in the published scientific literature [191]. In this work we used the information about ungulates (Order *Artiodactyla* and *Perissodactyla*), carnivores and primates. Based on the data obtained from the two databases, we constructed four bipartite networks (Parasite-Fish, Parasite-Ungulates, Parasite-Carnivores and Parasite-Primates) in which the first set of nodes is composed by parasites and the second by hosts, linked if the parasite interacts with the host. The best model found was:

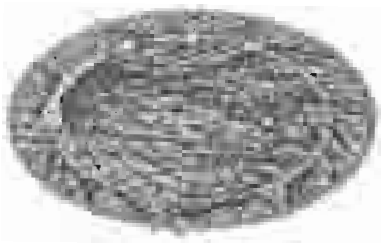
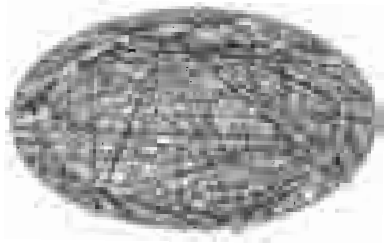
$$\begin{aligned} S(\text{parasite-host}) &= 2.4756 \cdot \pi_3(\text{p-h}) + 0.2114 \\ n &= 49,218 \quad \chi^2 = 15,804.87 \quad p < 0.001 \end{aligned}$$

Table 5. Classification Matrices

		Analysis			Cross-Validation		
		Percent	0	1	Percent	0	1
Drug-Target	0	76.01[a]	1,150	363	76.94[a]	367	110
	1	81.83[b]	131	590	79.17[b]	50	190
	Total	**77.89**[c]	0.25[d]	0.75[d]	**77.68**[c]	0.25[d]	0.75[d]
Parasite-Host	0	87.49	39,047	5,581	87.67	12,972	1,824
	1	100.00	0	4,590	100.00	0	1,521
	Total	**86.66**	0.5	0.5	**88.82**	0.5	0.5
Fasciolosis spreading	0	93.21	20,291	1,477	93.20	6,812	497
	1	72.01	890	2,290	73.47	281	778
	TOTAL	**90.51**	0.5	0.5	**90.70**	0.5	0.5
C. elegans metabolic network	0	98.81	17,020	205	98.91	5,645	62
	1	76.21	511	1,637	77.09	164	552
	TOTAL	**96.30**	0.5	0.5	**96.48**	0.5	0.5
Macaque brain	0	98.49	29,268	448	98.41	9,766	158
	1	73.30	1,322	3,630	71.21	475	1,175
	TOTAL	**94.89**	0.5	0.5	**94.53**	0.5	0.5
Financial law	0	82.77	15,698	3,267	82.62	5,239	1,102
	1	100.00	0	14,986	100.00	0	5,014
	TOTAL	**90.38**	0.5	0.5	**90.30**	0.5	0.5

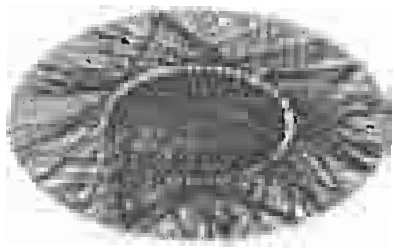
Rows: Observed classifications; Columns: Predicted classifications. [a] Specifity; [b] Sensitivity; [c] Accuracy. 0 = non connected, 1 = connected, [d] LDA *a priori* probabilities.

Table 6. Comparison of Observed *vs.* Reconstructed Interactions in Drug-Target Networks (Giant Components)

Drug-Target			
Observed Network	Network Descriptors		Reconstructed Network
	638	n	539
	404	nd	320
	234	nt	219
	802	m	671
	2.51	Ad	2.49
	0.00848	den	0.00957
	2,702,728	W	2,122,796
	17	D	18
	6.65	AD	7.32
			

[a] Network descriptors: Total number of connected nodes (n), number of connected drugs (nd), number of connected targets (nt), number of edges (m), average degree (Ad), density (den), Wiener index (W), diameter (D) and average distance (AD).
[b] The size of the drawn nodes is proportional to its degree. [c] Outer nodes: Drugs; Inner nodes: Targets.

Table 7. Comparison of Observed *vs.* Reconstructed Interactions in Parasite-Host Networks

Parasites-Fish			
Observed Network	Network Descriptors[a]		Reconstructed Network
	298	n	298
	239	np	239
	59	nh	59
	912	m	912
	6.12	Ad	6.12
	0.0647	den	0.0647
	298,626	W	298,626
	7	D	7
	3.37	AD	3.37