



UNIVERSIDADE DA CORUÑA

Facultad de Informática

Departamento de
Tecnologías de la Información y las
Comunicaciones

Técnicas de ingeniería informática e inteligencia artificial para clasificación: aplicaciones para el descubrimiento de fármacos y dianas moleculares

Tesis Doctoral

Directores

Alejandro Pazos Sierra

Humberto González Díaz

Doctorando

Cristian Robert Munteanu

A Coruña, Abril 2013

Dr. Alejandro Pazos Sierra, Catedrático de Universidad en el área de Ciencias de la Computación e Inteligencia Artificial, perteneciente al Departamento de Tecnologías de la Información y las Comunicaciones, Facultad de Informática, Universidade da Coruña

Y

Dr. Humberto González Díaz, Prof. Investigador Ikerbasque del Departamento de Química Orgánica II, Facultad de Ciencia y Tecnología, Universidad del País Vasco, UPV/EHU

HACEN CONSTAR QUE:

La memoria “**Técnicas de ingeniería informática e inteligencia artificial para clasificación: aplicaciones para el descubrimiento de fármacos y dianas moleculares**” ha sido realizada por **D. Cristian Robert Munteanu**, bajo nuestra dirección en el Departamento de Tecnologías de la Información y las Comunicaciones, y constituye la Tesis que presenta para optar al Grado de Doctor en Informática de la Universidade da Coruña.

A Coruña, 24 de Abril de 2013

Fdo: Alejandro Pazos Sierra

Fdo: Humberto González Díaz

A mi hijo, Tudor

Agradecimientos

Esta Tesis Doctoral ha sido realizada en el Tecnologías de la Información y las Comunicaciones, Facultad de Informática, Universidade da Coruña bajo la supervisión del Dr. Alejandro Pazos Sierra y Dr. Humberto González Díaz, a los cuales me gustaría agradecer la inestimable ayuda que me ha prestado.

También agradezco toda la ayuda que me han prestado los colaboradores de la Universidad de Santiago de Compostela, especialmente a Dr. Francisco Prado-Prado y para el soporte informático brindado por el grupo de Redes de Neuronas Artificiales y Sistemas Adaptativos (Universidade da Coruña), especialmente a Julián Dorado, Vanessa Aguiar-Pulido y Dr. Marcos Gestal Pose.

Quisiera extender estos agradecimientos también a los profesores que me formaron como científico, Matei y Florentina Ion, Hillebrand Mihaela, Domnina Razus y Berta Fernández Rodríguez.

Quiero hacer una mención especial a mi familia y a mis amigos sin cuyo esfuerzo y apoyo no habría sido posible que hubiese llegado hasta aquí.

¡Muchas gracias a todos!

Cristian R. Munteanu

Resumen

La búsqueda de nuevos fármacos y sus dianas moleculares tiene mucho interés en la industria farmacológica, con implicaciones en práctica clínica contra enfermedades complejas, especialmente contra los microbios y parásitos. Como la búsqueda experimental de la acción biológica de todos los fármacos posibles y de sus dianas es una actividad muy costosa y que implica mucho tiempo, surge la necesidad utilizar métodos teóricos para predecir los mejores candidatos.

La tesis aquí propuesta plantea el desarrollo de nuevas herramientas informáticas para el descubrimiento de fármacos y dianas moleculares, utilizando técnicas de ingeniería informática e inteligencia artificial. En consecuencia, la información estructural de las moléculas se codificó en los índices topológicos de los grafos moleculares, con la ayuda de nuevos programas informáticos específicos implementados por el autor de la tesis. Con estos índices, se buscaron modelos de clasificación capaces de predecir la actividad biológica de nuevas moléculas o la interacción entre moléculas. Los mejores modelos desarrollados se implementaron como herramientas informáticas “Web” con acceso libre para los científicos. Todos los resultados se publicaron en revistas internacionales con importante factor de impacto JCR.

Abstract

The search for new drugs and their molecular targets have an increased interest for the pharmaceutical industry with implications in clinical practice against complex diseases, especially against microbes and parasites. The experimental search of the biological activity of all possible drugs and their targets is very expensive and involves a lot of time. Therefore, it appears the necessity of theoretical methods to predict the best candidates.

The current thesis proposes the development of new tools for drug discovery and their molecular targets using software engineering and artificial intelligence techniques. Consequently, structural information was encoded in molecules topological indices of molecular graphs with the help of new specific computer programs. These indices are used to seek the classification models that can predict the biological activity of new molecules or the interaction of molecules (drugs / proteins). The best models were implemented as Web tools with free access to the scientific world. All results were published in international journals with JCR impact factor.

Resumo

A busca de novos fármacos e os seus obxectivos moleculares teñen moito interese na industria farmacéutica, con implicacións na practica clínica contra enfermidades complexas, nomeadamente contra os microbios e parasitos. Como a procura experimental da acción biolóxica de todas as drogas posibles e as súas dianas é moi custoso e implica moito tempo, xorde a necesidade de empregar métodos teóricos para prever os mellores candidatos.

A tese aquí proposta fai énfase no desenvolvemento de novas ferramentas para o descubrimento de drogas e dianas moleculares utilizando técnicas de enxeñería informática e intelixencia artificial. En consecuencia, a información estrutural das moléculas foi codificada en índices topolóxicas de grafos moleculares, coa axuda de novos programas informáticos específicos implementados por o autor da tese. Con estes índices, procuráronse novos modelos de clasificación que poidan predicir a actividade biolóxica de novas moléculas ou a interacción entre moléculas. Os mellores modelos acadados foron implementados en ferramentas “Web”, con acceso gratuíto para os científicos. Todos os resultados foron publicados en revistas internacionais con importante factor de impacto JCR.

1. INTRODUCCIÓN	1
1.1. Programas para parámetros de grafos moleculares	6
1.1.1. DRAGON.....	7
1.1.2. MoDesLab.....	7
1.1.3. TOMO-COMD.....	8
1.1.4. MARCH-INSIDE.....	9
1.1.5. E-Calc.....	9
1.1.6. CODESSA PRO.....	10
1.2. Modelos de inteligencia artificial para fármacos y dianas moleculares	12
1.2.1. Modelos de clasificación para compuestos anti-virales.....	12
1.2.2. Modelos de clasificación para compuestos anti-bacterianos.....	14
1.2.3. Modelos de clasificación para compuestos anti-parasitarios.....	15
1.2.4. Modelos de clasificación para compuestos anti-fúngicos.....	17
1.3. Herramientas online de clasificación molecular	20
1.4. Objetivos	24
2. RESULTADOS Y DISCUSIÓN	25
2.1. Nuevos programas de ordenador para los parámetros moleculares	26
2.1.1. MInD-Prot – Descriptores Markov para fármacos y proteínas.....	26
2.1.2. S2SNet – Índices topológicos del grafo tipo estrella.....	31
2.1.3. CULSPIN – Índices topológicos del grafo tipo espiral.....	42
2.2. Nuevos servidores online Bio-AIMS basados en técnicas de ingeniería informática e inteligencia artificial	57
2.2.1. Trypano-PPI – Interacciones proteína-proteína en <i>Trypanosoma</i>	59
2.2.2. Plasmod-PPI – Interacciones proteína-proteína en <i>Plasmodium</i>	62
2.2.3. ATCUNpred – Predicción de dianas proteicas con actividad ATCUN en parásitos	64
2.2.4. LIBPpred – Predicción de proteínas que interacciona con los lípidos.....	66
3. CONCLUSIONES	69
4. REFERENCIAS	70
5. PUBLICACIONES (ANEXOS)	

Publicaciones con S2SNet

Enrique Fernandez-Blanco, Vanessa Aguiar-Pulido, **Cristian R Munteanu**, Julian Dorado, *Random Forest Classification based on Star Graph Topological Indices for Antioxidant Proteins*, [Journal of Theoretical Biology](http://dx.doi.org/10.1007/s10440-013-0317-3) 317, 331-337 (2013) [<http://goo.gl/R5vV8>]

Publicaciones con grafos de tipo espiral

Vanessa Aguiar-Pulido, **Cristian Robert Munteanu**, José A Seoane, Enrique Fernández-Blanco, Lázaro G Pérez-Montoto, Humberto González-Díaz, Julian Dorado, *Naïve Bayes QSDR classification based on spiral-graph Shannon entropies for protein biomarkers in human colon cancer*, Molecular BioSystems 8, 1716-1722 (2012) [<http://goo.gl/JQQIE>]

Publicaciones para los servidores Web

Yamilet Rodriguez-Soca, **Cristian R. Munteanu**, Julián Dorado, Alejandro Pazos, Francisco J. Prado-Prado, and Humberto González-Díaz, *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(2), 1182–1190 (2010) [<http://goo.gl/nCgR9>]

Herramienta: <http://bio-aims.udc.es/TrypanoPPI.php>

Yamilet Rodriguez-Soca, **Cristian R. Munteanu**, Julian Dorado, Juan Rabuñal, Alejandro Pazos and Humberto González-Díaz, *Plasmod-PPI: a web-server predicting complex biopolymer targets in Plasmodium with entropy measures of protein-protein interactions*, Polymer 51(1), 264-273 (2010) [<http://goo.gl/hRhm9>]

Herramienta: <http://bio-aims.udc.es/PlasmodPPI.php>

Cristian R Munteanu, José M. Vázquez, Julián Dorado, Alejandro Pazos Sierra, Ángeles Sánchez-González, Francisco J. Prado-Prado and Humberto González-Díaz, *Complex Network Spectral Moments for ATCUN Motif DNA Cleavage: First Predictive Study on Proteins of Human Pathogen Parasites*, Journal of Proteome Research 8(11), 5219–5228 (2009) [<http://goo.gl/u7Thq>]

Herramienta: <http://bio-aims.udc.es/ATCUNPred.php>

Humberto González-Díaz, **Cristian R. Munteanu**, Lucian Postelnicu, Francisco Prado-Prado, Marcos Gestal and Alejandro Pazos, *LIBP-Pred: web server for lipid binding proteins using structural network parameters; PDB mining of human cancer biomarkers and drug targets in parasites and bacteria*, Molecular BioSystems 8, 851-862 (2012) [<http://goo.gl/cTNcP>]

Herramienta: <http://bio-aims.udc.es/LIBPpred.php>

1. INTRODUCCIÓN

Desde cuando se ha manifestado el interés médico en los microbios y parásitos, los científicos intentaron encontrar los métodos más eficaces para combatir los efectos negativos en la salud de las personas. En esta lucha, los organismos dianas están aprendiendo continuamente a desarrollar resistencia contra los fármacos actuales y a adaptarse a nuevas condiciones del entorno. Por ello, se necesitan métodos rápidos, accesibles y baratos para descubrir nuevos fármacos y dianas moleculares contra los microbios y parásitos. Los métodos teóricos son una opción excelente para encontrar más rápido y con menos recursos materiales y humanos nuevos tratamientos para mejorar la calidad de vida de las personas.

La tesis actual propone el desarrollo de nuevas aplicaciones y programas informáticos para el descubrimiento de fármacos y dianas moleculares utilizando técnicas de ingeniería informática e inteligencia artificial para clasificación. Se desarrollan modelos teóricos basados en la teoría de las redes complejas o del grafo y en las técnicas de las relaciones cuantitativas estructura-actividad o propiedad (QSAR/QSPR) y la implementación de los mejores modelos en herramientas gratis online, accesible desde cualquier parte del mundo. Para poder desarrollar este tipo de solución, se necesitan estudios interdisciplinarios con conocimientos y métodos de los siguientes campos: Química Farmacéutica para comprender la actividad de los fármacos, Microbiología y Parasitología para encontrar la mejor forma de luchar contra diversas patologías, Bioinformática para manipular la información biológica, Matemáticas Aplicadas con la teoría de los grafos y de las redes complejas para caracterizar numéricamente los fármacos y sus dianas moleculares en microbios y parásitos, Inteligencia Artificial y Estadística para encontrar los modelos teóricos que pueden predecir nuevos fármacos y sus dianas e Informática con técnica de programación para crear las aplicaciones que pueden generar descriptores moleculares y para implementar los modelos de predicción en herramientas online únicas en todo el mundo científico.

Las QSAR/QSPR, acrónimo del inglés *Quantitative Structure-Activity/Property Relationships*, han sido ampliamente utilizadas para diferentes tipos de problemas en Química Médica y otras Ciencias Biológicas. Sin embargo, las aplicaciones de los modelos QSAR se han limitado al estudio de pequeñas moléculas en el pasado. En este contexto, muchos autores utilizan grafos moleculares de átomos (nodos) conectados por enlaces químicos (aristas) para representar y caracterizar numéricamente la estructura molecular. Sin embargo, más

recientemente, han aparecido muchos modelos QSAR/QSPR con aplicaciones a situaciones más generales. Por ejemplo, los nuevos modelos pueden aplicarse para predecir la función de una proteína con una secuencia o una estructura determinada en 3D, la función de una estructura secundaria del ARN, las interacciones de los fármacos específicos con múltiples dianas (como proteínas) presentes en el proteoma de un organismo o varios organismos infecciosos/parasitarios [1, 2]. En este sentido se han publicado diferentes trabajos para discutir tanto las aplicaciones clásicas del QSAR, como también otras nuevas en distintas áreas/revistas: 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].

En todos estos trabajos de revisión se puede observar que la teoría de grafos y redes complejas se está expandiendo a diferentes niveles de organización de la materia tales como las redes del genoma, las redes de interacción proteína-proteína, redes huésped-parásito, redes lingüísticas, redes sociales [49-54], redes electro-energéticas e Internet [55]. Una red es un conjunto de elementos, generalmente llamados nodos, con conexiones entre ellos (aristas). Los nodos pueden ser átomos, moléculas, proteínas, ácidos nucleicos, fármacos, células, organismos, parásitos, personas, leyes, ordenadores o cualquier otro componente de un sistema real. Las aristas son las relaciones entre los nodos, como los enlaces químicos, las interacciones físicas, las vías metabólicas, la acción farmacológica, la recurrencia de la ley o el comportamiento social [54]. Para el estudio cuantitativo, las redes complejas se pueden caracterizar numéricamente por parámetros únicos de la red habitualmente conocidos como índices topológicos (TIs). Los TIs de redes conocidas (moleculares o no) se utilizan como entradas en el análisis estadístico para construir modelos tipo QSAR/QSPR. En este sentido se han desarrollado distintos programas para el cálculo de estos parámetros.

En consecuencia, se pueden definir los siguientes elementos en la teoría de las redes complejas que se utilizarán a lo largo de toda la tesis:

- ❖ red - un grupo interconectado o sistema de elementos que comparte información;
- ❖ grafo - representación simbólica de una red y de su conectividad; implica una abstracción de la realidad por la que se puede simplificar como un conjunto de nodos (vértices) conectados por líneas (aristas) que representan las relaciones/propiedades comunes;
- ❖ índices topológicos - cualquier parámetro numérico invariante de un grafo que caracteriza su topología/geometría/estructura; codifican la información sobre las funciones de la red real.

El esquema general del trabajo con técnicas QSAR y la teoría de las redes complejas está presentado en **Figura 1**:

- las moléculas de proteínas o fármacos (redes reales de aminoácidos y átomos) están transformados en grafos específicos: en el caso de las proteínas, los nodos son los carbonos alpha de los aminoácidos desde la estructura 3D y en el caso de los fármacos los nodos son todos los átomos de la fórmula química (códigos SMILES); para eso se desarrollaron tres programas informáticos que pueden calcular descriptores moleculares utilizando diferentes tipos de grafos: MInD-Prot, S2SNet y CULSPIN;
- estos grafos se caracterizan por unos índices topológicos/descriptores moleculares que se basan en matrices de conectividad, distancias entre nodos, grados de enlace de los nodos y probabilidades de transición;
- estos números específicos para cada molécula con una actividad biológica específica se pueden utilizar para crear modelos de clasificación QSAR mediante análisis discriminante general, redes neuronales artificiales, aprendizaje automático, computación evolutiva, etc.; con estos modelos se pueden evaluar nuevos fármacos y dianas proteicas para una función biológica específica;
- los mejores modelos se implementan en una colección de cuatro herramientas online en el servidor Bio-AIMS (<http://bio-aims.udc.es>): Trypano-PPI para el estudio de las interacciones proteína-proteína en *Trypanosoma*, Plasmod-PPI para las interacciones proteína-proteína en *Plasmodium*, ATCUNpred para la actividad ATCUN de las proteínas, con aplicación en parásitos como *Trypanosoma*, *Plasmodium*, *Leishmania*, o *Toxoplasma* y LIBPpred para la predicción de proteínas que interacciona con los lípidos en *Shigella flexneri*, *Plasmodium berghei* y *Cryptosporidium parvum*;
- las herramientas se pueden utilizar para el descubrimiento de nuevos fármacos y sus dianas proteicas, interacciones proteínas – proteínas o nuevas proteínas con una actividad específica.

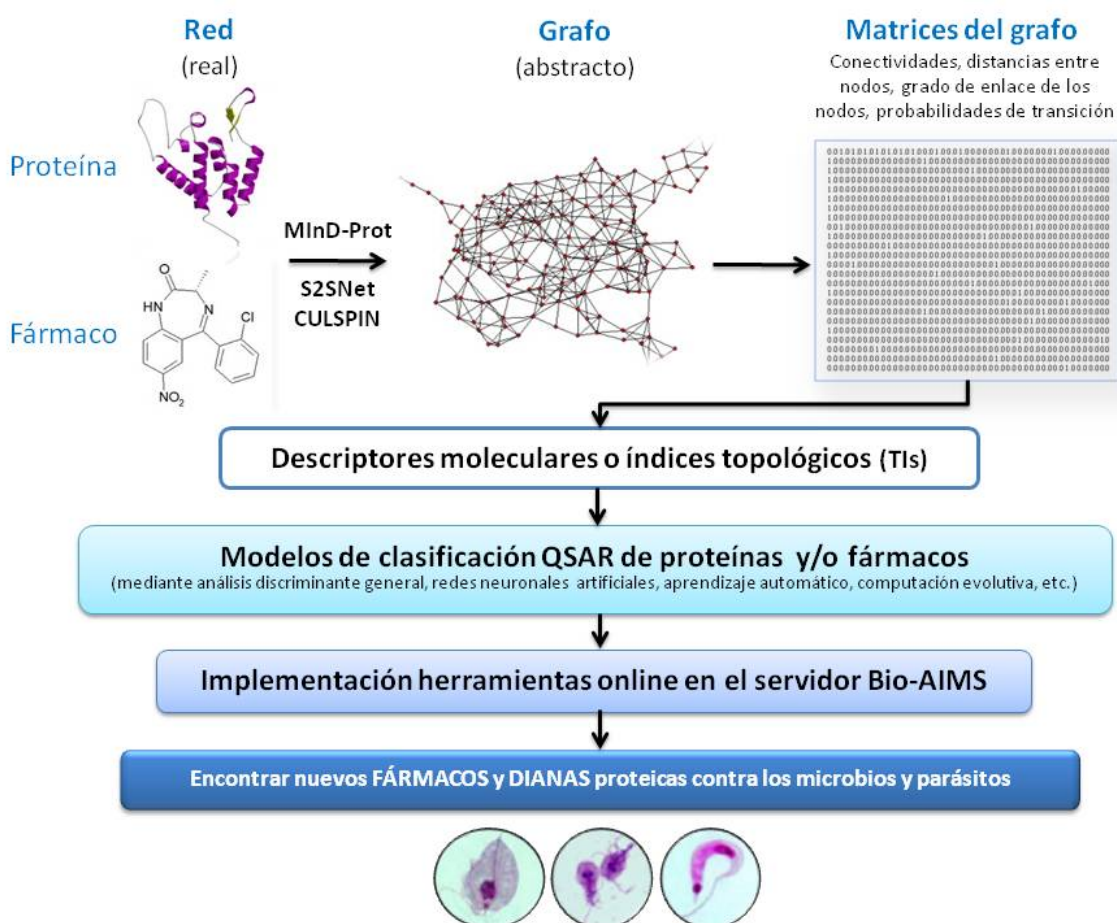


Figura 1: La esquema general del trabajo con técnicas QSAR y la teoría de las redes complejas, el descubrimiento de fármacos y dianas moleculares

La sección **INTRODUCCIÓN** comienza describiendo los programas informáticos existentes para el cálculo de los descriptores moleculares/índices topológicos tales como DRAGON, MoDesLab, TOMO-COMD, MARCH-INSIDE, E-Calc y CODESSA PRO. La misma sección continúa con la presentación de los modelos existentes de tipo QSAR/QSPR para compuestos anti-virales, anti-bacterianos, anti-parasitarios y anti-fúngicos. A continuación, se presentarán unos ejemplos de herramientas Web basados en modelos de inteligencia artificial.

Después de la revelación de los objetivos de esta tesis, comienza la segunda sección, la de los **RESULTADOS Y DISCUSION** dividida a su vez en tres partes: nuevos programas informáticos para el cálculo de los descriptores moleculares, nuevas herramientas online en el Bio-AIMS que se basan en modelos de clasificación QSAR y la presentación de trabajos de revisión y los capítulos de libros dedicados a las aplicaciones de grafos en las ciencias Bio.

Todas las partes de esta sección contienen el sumario de las publicaciones correspondientes.

La tesis continúa con las **CONCLUSIONES**, las **REFERENCIAS** en el texto hasta ese punto y una sección que incluye las seis **PUBLICACIONES (ANEXOS)** con índice de impacto JCR, en el lenguaje original que corresponden a los sumarios presentados anteriormente en la parte de los **RESULTADOS Y DISCUSIONES**.

1.1. Programas para parámetros de grafos moleculares

Muchos fenómenos pueden ser modelados como una red compleja. Por eso, la teoría de redes se puede utilizar en los estudios sobre el descubrimiento de fármacos, las vías metabólicas, enfermedades, búsqueda de dianas moleculares, interacciones entre macromoléculas etc. En esta tesis vamos a centrarnos tanto en los sistemas moleculares tales como los fármacos y las proteínas, como también en sus dianas moleculares.

Los descriptores moleculares juegan un papel fundamental en los estudios QSPR/QSAR. En esta sección vamos a presentar algunos programas que se utilizan para el cálculo de descriptores moleculares (tanto TIs como otros) [56]: DRAGON, MoDesLab, TOMO-COMD, MARCH-INSIDE, E-Calc y CODESSA PRO.

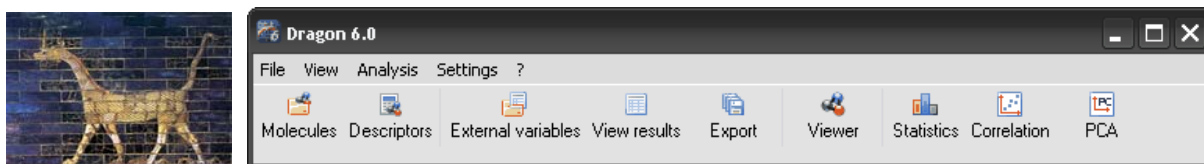


Figura 2: Interfaz gráfica de la aplicación **Dragon 6**

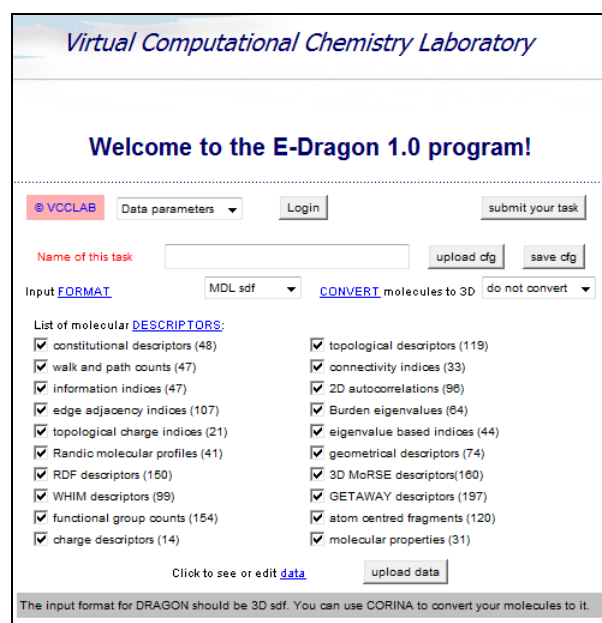


Figura 3: Interfaz de la versión online **E-Dragon 1.0**

1.1.1. DRAGON

El programa DRAGON (http://www.talete.mi.it/products/dragon_description.htm) ha sido concebido para proporcionar al usuario una variedad de descriptores moleculares (incluyendo la mayoría de los TIs conocidos) derivados de las diferentes representaciones moleculares (**Figura 2**). El primer lanzamiento de DRAGON fue desarrollado en 1994 por el Grupo Milano Chemometrics con el nombre WHIM/3D QSAR. Sucesivamente, se han incorporado una gran cantidad de descriptores, dando lugar a un nuevo software, que en 1997 proporcionó unos 600 descriptores y se publicó con el nombre de DRAGON [57]. En la actualidad, DRAGON (v. 6.0) permite el cálculo de 4855 descriptores moleculares divididos en 29 tipos y es administrado por Talete SRL, una marca comercial. E-DRAGON (v. 1.0) (<http://www.vcclab.org/lab/edragon/>) es la versión online de DRAGON (v. 5.4) (**Figura 3**). Es gratuito y permite el cálculo de más de 4885 descriptores moleculares que se dividen en 20 bloques lógicos [58]. E-Dragón ha sido desarrollado como resultado de la colaboración entre el Dr. Tetko, el profesor Todeschini y los equipos del Prof. de Gasteiger. Algunos ejemplos en la literatura sobre el uso de este software son [59-61].

1.1.2. MoDesLab

MoDesLab (<http://www.modeslab.com/>), ha sido desarrollado por E. Estrada y Gutiérrez Y. y fue lanzado por primera vez en 2002 (**Figura 4**). Actualmente podemos encontrar la versión 1.5, lanzada en 2004. Proporciona todas las herramientas necesarias para llevar a cabo estudios QSAR, a partir de la entrada de un gran número de moléculas para el cálculo de descriptores moleculares (por ejemplo, Kier y Hall, índices Kappa, los índices de Balaban, los descriptores de Abraham y descriptores sub-estructurales propios del TOPS-MODE). También proporciona una manera muy útil para definir las propiedades de los átomos, enlaces y fragmentos así como permite introducir las estructuras moleculares en el lenguaje SMILES para el uso de estas propiedades en el cálculo de los descriptores moleculares [62-64].

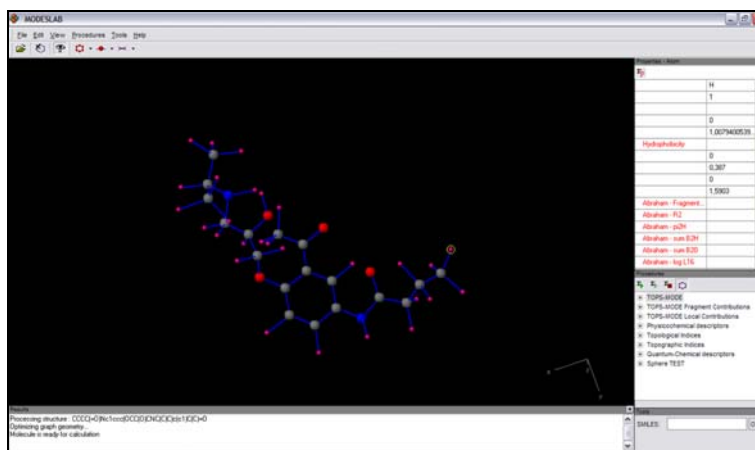


Figura 4: Interfaz del MoDesLab

1.1.3. TOMO-COMD

En 2002 Y. Marrero-Ponce y Romero V. han lanzado la versión 1.0 de TOMOCOMD (**Figura 5**). Se compone de cuatro subprogramas y cada uno de ellos permite tanto la edición de las estructuras (modo de dibujo) como el cálculo de descriptores moleculares 2D/3D (modo de cálculo). El software calcula distintos tipos de TIs a partir de formas algebraicas: tales como la cuadrática $qk(w)$, la lineal $fk(w)$ y la bi-lineal $bk(w, v)$ [65]. En un trabajo reciente de revisión se han discutido muchas aplicaciones de TOMOCOMD en estudios QSPR/QSAR de fármacos anti-parasitarios [35].

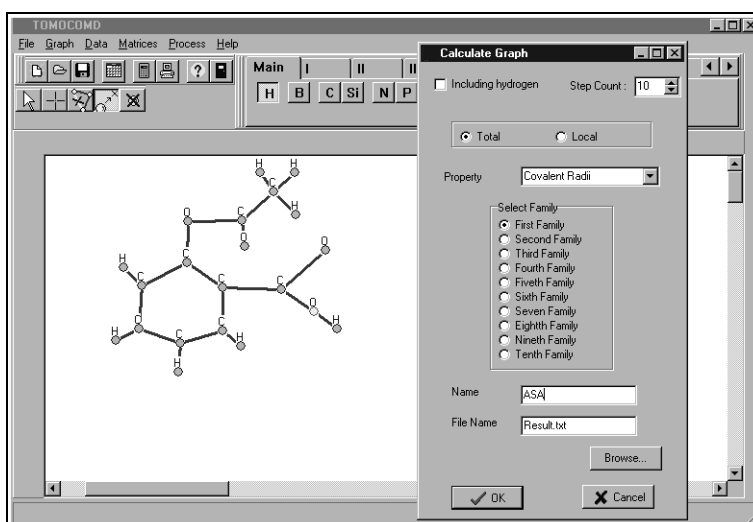


Figura 5: Interfaz del TOMOCOMD

1.1.4. MARCH-INSIDE

MARCH-INSIDE es un método de cálculo simple, pero eficaz para el estudio QSAR en la química medicinal, desarrollado por González-Díaz *et al.* (**Figura 6**). Se utiliza la teoría de las cadenas de Markov para generar parámetros que describen numéricamente la estructura química de los fármacos y sus dianas moleculares. En trabajos de revisión recientes podemos encontrar ejemplos de la utilización de este programa en la predicción de agentes antimicrobianos y anti-parasitarios, así como sus dianas moleculares [10, 35, 66].

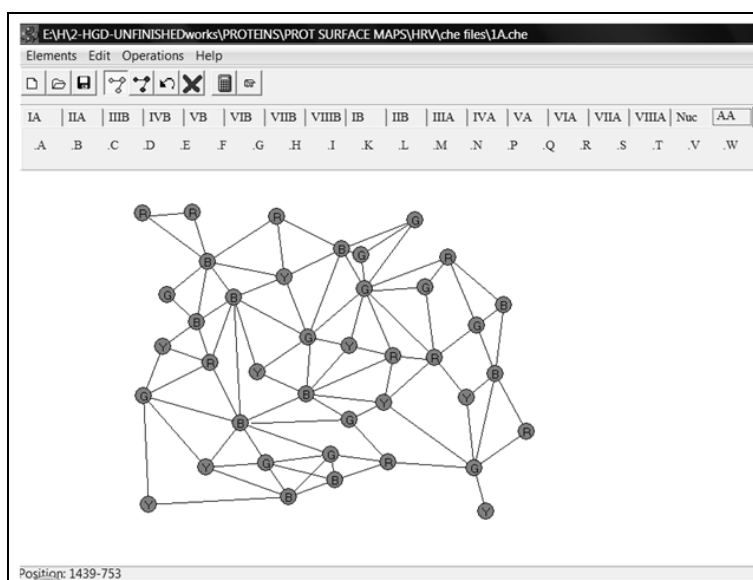


Figura 6: Interfaz gráfica de la aplicación MARCH-INSIDE

1.1.5. E-Calc

E-Calc (v.1.1/1999) es una utilidad que calcula índices del Estado Electrotopológico (E-valores) de las moléculas, incluyendo el estado electrotopológico (E-Estado) y el E-Estado de hidrógeno (HE-Estado), los valores de los átomos individuales, así como los índices del átomo (**Figura 7**). Estos cálculos ayudan a entender el desarrollo, uso e interpretación de los valores

del E-Estado como una representación de la estructura molecular. Las partes de cómputo de este programa se han tomado de Molconn-Z y de SciQSAR 2D [67].

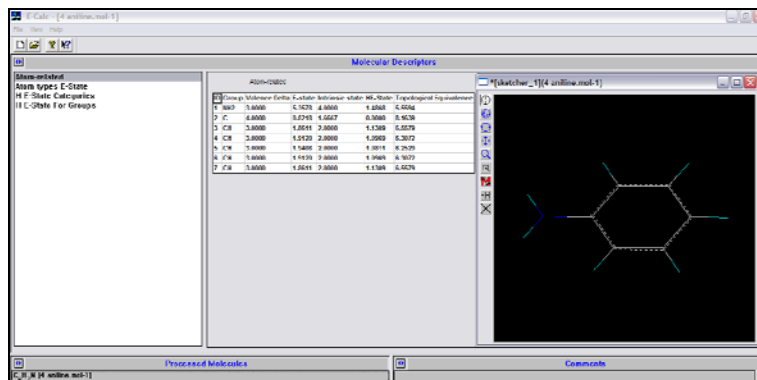


Figura 7: Interfaz del E-Calc

1.1.6. CODESSA PRO

CODESSA PRO, Comprehensive Descriptors for Structural and Statistical Analysis (análisis estructural y estadística para descriptores, <http://www.codessa-pro.com/>) es un programa diseñado por Alan R. Katritzky, Karelson Mati y Petrukhin Ruslan y desarrollado del año 2001 al 2005 (**Figura 8**). El manual del usuario (<http://www.codessa-pro.com/manuals/manual.htm>) especifica que está diseñado para el desarrollo de las relaciones cuantitativas tipo QSAR/QSPR mediante la integración de todas las medidas matemáticas y herramientas computacionales necesarias para: (i) calcular una gran variedad de descriptores moleculares utilizando la estructura geométrica 3D y/o la función de onda mecanocuántica de los compuestos químicos, (ii) el desarrollo (de varios) modelos QSPR lineales y no lineales para propiedades químicas y físicas o para la actividad biológica de los compuestos químicos, (iii) llevar a cabo un análisis de agrupamiento en clústeres de datos experimentales y descriptores moleculares, (iv) interpretar los modelos desarrollados, y (v) predecir los valores de propiedad de cualquier compuesto químico con una estructura molecular conocida. CODESSA PRO incluye 116 descriptores moleculares divididos en 8 grupos: constitucionales, topológicos y geométricos, CPSA electrostáticos, cuánticos, químicos, relacionados con las orbitales moleculares y la termodinámica. Algunos ejemplos del uso de este programa de investigación están en [68-71].

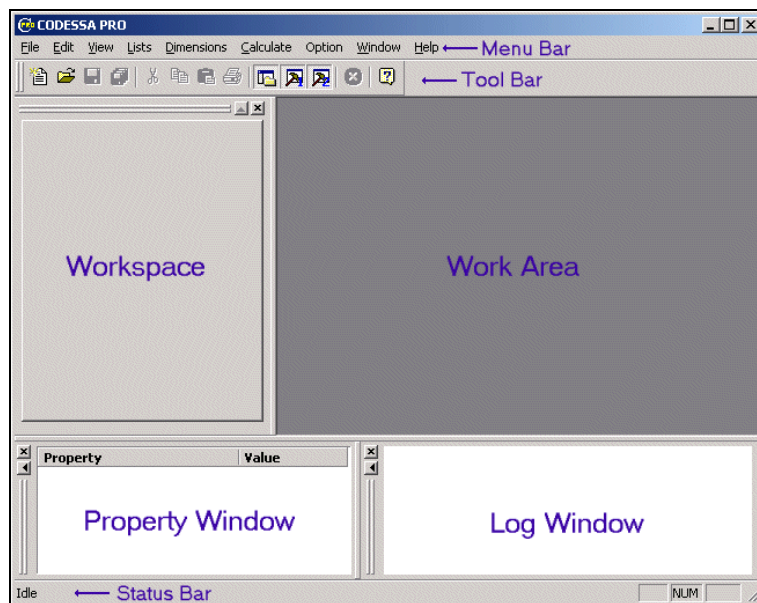


Figura 8: La interfaz visual de la aplicación **CONDESSA PRO**

1.2. Modelos de inteligencia artificial para fármacos y dianas moleculares

La búsqueda experimental de nuevos fármacos y dianas moleculares para luchar contra los microbios y parásitos implica un esfuerzo financiero y humano. Por esta razón, los científicos necesitan unos métodos teóricos extremadamente rápidos y baratos para predecir actividades biológicas de nuevos posibles fármacos o proponer posibles dianas moleculares. Por eso, se utilizan como modalidad inicial de screening los modelos QSAR que pueden establecer una relación cuantitativa entre la estructura química de los fármacos/dianas moleculares y la actividad biológica/capacidad específica de interaccionar. Una limitación de casi todos los modelos QSAR/QSPR es que predicen la actividad biológica de los medicamentos sólo para un sistema biológico (organismo, diana etc.). La solución viene con el desarrollo de modelos múltiples tareas QSAR/QSPR (mt-QSAR/mt-QSPR) para predecir la actividad de los fármacos/propiedades contra diferentes sistemas biológicos. Estos mt-QSAR/mt-QSPRs ofrecen también una buena oportunidad para la construcción de redes complejas que se pueden utilizar para explorar grandes y complejas bases de datos de medicamentos-sistemas biológicos. En esta sección vamos a revisar algunos de los modelos (mt-)QSAR/QSPR propuestos en la literatura y las redes de deriva de estos estudios.

1.2.1. Modelos de clasificación para compuestos anti-virales

Prado-Prado *et al.* [72] han utilizado la teoría de la cadena de Markov para calcular nuevos momentos espectrales para múltiples dianas con el fin de ajustar un modelo mt-QSAR para medicamentos activos contra 40 especies virales. El modelo se basa en 500 medicamentos (incluidos compuestos activos e inactivos) analizados como agentes antivirales en la literatura reciente; no todos los medicamentos fueron evaluados contra todos los virus, sólo aquellos con valores experimentales. La base de datos también contiene 207 compuestos conocidos (que no son tan recientes como los anteriores).

Estos compuestos han sido presentados en el Índice de Merck con otras actividades, que no incluyen la acción antiviral contra cualquier especie de virus, y han sido utilizados como compuestos inactivos. El Análisis Discriminante Lineal (LDA) se ha empleado para clasificar todos estos medicamentos en dos clases, de compuestos activos o inactivos contra las diferentes especies virales analizadas. El modelo clasificó correctamente 5129 de los 5594 compuestos

inactivos (sensibilidad = 91,69%) y 412 de los 422 compuestos activos (especificidad = 97,63%). La ecuación del modelo es la siguiente:

$$\begin{aligned}
 Actv = & -0.95 \cdot^0 \mu_s(\text{H-Het}) + 1.50 \cdot^2 \mu_s(\text{H-Het}) - 3.23 \cdot^0 \mu_s(\text{C}_{\text{uns}}) \\
 & - 4.02 \cdot^0 \mu_s(\text{C}_{\text{sat}}) - 0.47 \cdot^1 \mu_s(\text{T}) + 10.34 \cdot^0 \mu_s(\text{T}) + 0.74 \cdot^5 \mu_s(\text{X}) \\
 & - 8.88 \\
 \lambda = & 0,51; \quad \chi^2 = 4024.83; \quad p < 0.001
 \end{aligned} \tag{1}$$

donde λ es la estadística de Wilk; χ^2 chi cuadrado y p el nivel de error. En la ecuación, $^k \mu_s$ es el momento espectral de una cierta especie después de k etapas. Se ha calculado para el total (T) de los átomos en la molécula o para asociaciones específicas de átomos. Estas asociaciones son átomos con una característica común: H-Het = hidrógeno unido a heteroátomos, C_{uns} = átomos de carbono insaturados, C_{sat} = átomos de carbono saturados, X = átomos de halógeno.

Prado-Prado *et al.* [73] han utilizado el LDA para ajustar un modelo mt-QSAR que ha clasificado 600 medicamentos como activos o inactivos contra 41 especies diferentes de virus analizadas. El modelo ha clasificado correctamente 143 de los 169 compuestos antivirales activos (especificidad = 84,62%) y 119 de los 139 compuestos inactivos (sensibilidad = 85,61%). La precisión en los datos de entrenamiento fue del 85,1% (262 de los 308 casos). Por otra parte, la validación del modelo se ha llevado a cabo utilizando la serie de predicción externa, obteniendo una precisión de validación cruzada de 90,7% (466 de los 514 compuestos). Para ilustrar el funcionamiento del modelo en la práctica, se desarrolló un *screening* virtual que reconoce como activos 102 de los 110 (92,7%) compuestos antivirales que no se utilizan en las series de entrenamiento o de predicción. La ecuación del modelo es la siguiente:

$$\begin{aligned}
 Actv = & 1.90 \cdot^0 C_s(\text{C}_{\text{sat}}) - 1.64 \cdot^0 C_s(\text{C}_{\text{uns}}) + 1.02 \cdot^2 C_s(\text{C}_{\text{uns}}) \\
 & + 1.10 \cdot^5 C_s(\text{C}_s) + 0.73 \cdot^1 C_s(\text{X}) + 1.08 \cdot^1 C_s(\text{Het}) \\
 & + 1.07 \cdot^0 C_s(\text{H-Het}) - 0.75 \cdot^4 C_s(\text{H-Het}) + 0.08 \\
 \lambda = & 0.47; \quad Rc = 0.726; \quad p < 0.001
 \end{aligned} \tag{2}$$

donde λ es la estadística de Wilk, Rc es la correlación canónica y p el nivel de error. En la ecuación $^k C_s$ es el índice molecular de una cierta especie después de k etapas. Se ha calculado para el total (T) de los átomos en la molécula o para asociaciones específicas de átomos presentadas en la ecuación anterior.

1.2.2. Modelos de clasificación para compuestos anti-bacterianos

Prado-Prado *et al.* [74] han desarrollado un modelo de Markov para describir la actividad biológica de más de 70 medicamentos de la literatura contra 96 especies de bacterias. Se ha aplicado el LDA para clasificar los medicamentos como activos o inactivos contra diferentes especies bacterianas analizadas. El modelo clasificó correctamente 199 de los 237 compuestos activos (83,9%) y 168 de los 200 compuestos inactivos (84%). La predictibilidad total en el grupo de entrenamiento fue del 84% (367 de los 437 casos). La validación del modelo se llevó a cabo utilizando la serie de predicción externa, clasificándose correctamente 202 de los 243 (83,13%) casos.

Con el fin de mostrar cómo funciona el modelo en la práctica, se llevó a cabo un *screening* virtual, el modelo reconociendo como activos 480 de los 568 (84,5%) compuestos antibacterianos que no se utilizaron en las series de entrenamiento o predicción. La ecuación del modelo es la siguiente:

$$\begin{aligned} Actv = & -1.12 \cdot C_s^1(T) + 1.34 \cdot C_s^3(T) + 1.84 \cdot C_s^0(C_{sat}) \\ & - 0.90 \cdot C_s^0(C_{uns}) + 0.88 \cdot C_s^5(X) - 1.27 \cdot C_s^0(H - Het) \\ & - 0.90 \cdot C_s^2(H - Het) + 0.698 \\ & \lambda = 0.49; \quad Rc = 0.715; \quad p < 0.001 \end{aligned} \quad (3)$$

donde λ es la estadística de Wilk, Rc es la correlación canónica y p el nivel de error. En la ecuación $^k C_s$ es el índice molecular de una cierta especie después de k etapas. Se ha calculado para el total (T) de átomos en la molécula o para asociaciones específicas de átomos.

Otro modelo, propuesto por Prado-Prado *et al.* [75], clasificó correctamente 202 de los 241 compuestos activos (83,8%) y 169 de los 200 casos inactivos (84,5%). La predictibilidad total en la serie de entrenamiento fue 84,13% (371 de los 441 casos). La validación del modelo se llevó a cabo utilizando la serie de predicción externa, clasificándose correctamente 197 de los 221 (89,4%) casos. La ecuación del modelo es la siguiente:

$$\begin{aligned} Actv = & -3.5 \cdot \pi_1(C_{sat}) + 3 \cdot \pi_0(C_{sat}) + 1.76 \cdot \pi_2(C_{uns}) - 1.77 \cdot \pi_3(Het) \\ & + 2.54 \cdot \pi_5(H - Het) + 2.4 \cdot \pi_3(Het - Het) - 5.42 \cdot \pi_2(H - Het) \\ & + 0.74 \\ & \lambda = 0.49; \quad Rc = 0.718; \quad p < 0.001 \end{aligned} \quad (4)$$

donde λ es la estadística de Wilk, Rc el índice canónico y p el nivel de error. En la ecuación, π_k es el momento espectral de una cierta especie después de k etapas. Se ha calculado para el total (T) de átomos en la molécula o para asociaciones específicas de átomos.

Los resultados de este modelo QSAR fueron utilizados como entradas para la construcción de una red. Esta red observada tiene 1242 nodos (medicamentos y bacterias), 772736 aristas (pares medicamento-bacteria) con una actividad similar. La red prevista tiene 1031 nodos y 641377 aristas. Después de una comparación de arista a arista, se ha demostrado que la red prevista es significativamente similar a la observada, y ambas tienen una distribución más cercana al exponencial que al normal.

1.2.3. Modelos de clasificación para compuestos anti-parasitarios

Prado-Prado *et al.* [76] han propuesto un mt-QSAR para más de 500 fármacos analizados en la literatura contra diferentes parásitos. Los datos fueron procesados por el LDA, clasificando los medicamentos como activos o inactivos contra diferentes especies de parásitos analizadas. El modelo clasificó correctamente 212 de los 244 (87,0%) casos de la serie de entrenamiento y 207 de los 243 compuestos (85,4%) de la serie de validación externa. Con el fin de ilustrar el funcionamiento de las QSAR para la selección de medicamentos activos se llevó a cabo un *screening* virtual adicional de compuestos antiparasitarios que no se utilizaron en las series de entrenamiento o predicción. El modelo reconoció 97 de 114 (85,1%) de ellos. La ecuación del modelo es la siguiente:

$$\begin{aligned}
 Actv = & 4.15 \times 10^{-14} \cdot {}^1C_s(T) + 8.9 \times 10^{-14} \cdot {}^0C_s(C_{sat}) \\
 & - 1.5 \times 10^{-13} \cdot {}^0C_s(C_{uns}) + 4.7 \times 10^{-7} \cdot {}^5C_s(C_{uns}) \\
 & + 2 \times 10^{-7} \cdot {}^0C_s(Het) - 7.9 \times 10^{-7} \cdot {}^4C_s(H-Het) - 0.72 \\
 Rc = & 0,75 \quad \lambda = 0.434; \quad F = 51,44; \quad p < 0.001
 \end{aligned}
 \tag{5}$$

donde Rc es el coeficiente de correlación canónica, λ es la estadística de Wilk, F es la relación de Fisher y p el nivel de error. En esta ecuación kC_s es el índice molecular de una cierta especie después de k etapas. Se ha calculado para el total (T) de átomos en la molécula o para asociaciones específicas de átomos. Estas asociaciones son átomos con una característica común: Het: heteroátomos, H-Het = hidrógeno unido a heteroátomos, C_{uns} = átomos de carbono insaturados, C_{sat} = átomos de carbono saturados.

Prado-Prado *et al.* [77] desarrollaron un modelo mt-QSAR para más de 700 medicamentos analizados en la literatura contra diferentes parásitos (medicamentos antiparasitarios de predicción). Los datos fueron procesados por el LDA y el modelo clasificó correctamente un 93,62% (1160 de los 1239 casos) en entrenamiento. La validación del modelo se llevó a cabo utilizando la serie de predicción externa, clasificándose correctamente 573 de los 607 (94,4%) casos. La ecuación del modelo es la siguiente:

$$\begin{aligned}
 Actv = & -3.86 \cdot \pi_1(s, C_{sat}) - 3.71 \cdot \pi_1(s, C_{sp\&sp2}) - 53.55 \cdot \pi_1(s, X) \\
 & + 50.92 \cdot \pi_3(s, X) - 2.62 \cdot \pi_1(s, H - Het) + 3.12 \cdot \pi_s(s, H - Het) \\
 & - 2.37 \\
 Rc = & 0.73; \quad \lambda = 0.46; \quad p < 0.001
 \end{aligned}
 \tag{6}$$

donde Rc es el coeficiente de correlación canónica, λ es la estadística de Wilk y p es el nivel de error. En esta ecuación, las probabilidades absolutas ${}^A\pi_k$ calculadas se refieren a: ${}^A\pi_{0,1}(s, C_{sp\&sp2})$: todos los átomos de carbono insaturados (átomos sp y sp2) y todos los átomos colocados a una distancia $d = 5$ de ellos. ${}^A\pi_1(s, C_{sat})$: todos los átomos de carbono saturados. ${}^A\pi_1(s, X)$: todos los átomos de halógeno. ${}^A\pi_0(s, H-Het)$: todos los átomos de hidrógeno unidos a un heteroátomo (N, O, o S).

Prado-Prado *et al.* [78] han utilizado la teoría de las Cadenas de Markov para calcular nuevos momentos espectrales para múltiples dianas con el fin de ajustar un modelo mt-QSAR para 500 medicamentos analizados en la literatura contra 16 especies de parásitos y otros 207 fármacos no analizados en la literatura. Los datos fueron procesados por el LDA, clasificando los medicamentos como activos o inactivos contra diferentes especies de parásitos analizadas. El modelo clasificó correctamente 311 de los 358 compuestos activos (86,9%) y 2328 de los 2577 compuestos inactivos (90,3%) en las series de entrenamiento. El rendimiento total de entrenamiento fue del 89,9%.

La validación del modelo se llevó a cabo mediante series de predicción externa. En estas series el modelo clasificó correctamente 157 de los 190 (82,6%) compuestos antiparasitarios y 1151 de los 1277 compuestos inactivos (90,1%). El rendimiento total de predictibilidad fue del 89,2%. Además, cuatro tipos de Redes Neuronales Artificiales (ANNs) no lineales fueron desarrolladas y comparadas con el modelo mt-QSAR. El modelo mejorado de ANN tuvo un rendimiento total de entrenamiento del 87%. La ecuación del modelo es la siguiente:

$$\begin{aligned}
Actv = & 1.49 \cdot^1 \mu_s(C_{uns}) + 1.12 \cdot^5 \mu_s(C_{uns}) + 1.92 \cdot^3 \mu_s(C_{sat}) \\
& + 0.53 \cdot^4 \mu_s(X) + 1.71 \cdot^1 \mu_s(H-Het) - 0.97 \cdot^2 \mu_s(H-Het) \\
& - 5.21 \\
\lambda = & 0.52 \quad \chi^2 = 1904.6; \quad p < 0.001
\end{aligned}
\tag{7}$$

El coeficiente λ es la estadística de Wilk; estadística de la discriminación total, χ^2 es el de chi-cuadrado y p es el nivel de error. En esta ecuación, μ_s se ha calculado para el total (T) de átomos en la molécula o para asociaciones específicas de átomos. Estas asociaciones son átomos con una característica común: H-Het: hidrógeno unido a heteroátomos, C_{uns} : átomos de carbono insaturados, C_{sat} : átomos de carbono saturados, X: átomos de halógeno.

1.2.4. Modelos de clasificación para compuestos anti-fúngicos

González-Díaz *et al.* [79] desarrollaron un modelo unificado de Markov para describir con una sola ecuación lineal la actividad biológica de 74 medicamentos analizados en la literatura contra algunas de las especies de hongos seleccionadas de una lista de 87 especies (491 casos en total). Los datos fueron procesados por el LDA, clasificando los medicamentos como activos o inactivos contra diferentes especies de hongos analizadas. El modelo clasificó correctamente 338 de los 368 compuestos activos (91,85%) y 89 de los 123 compuestos inactivos (72,36%). La predictibilidad total para el entrenamiento fue del 86,97% (427 de los 491 compuestos).

La validación del modelo se llevó a cabo mediante el método *leave-species-out* (LSO). Después de eliminar paso a paso todos los medicamentos analizados contra una especie, los autores registraron un porcentaje de buena clasificación de los compuestos *leave-species-out* (previsibilidad LSO). Además, se tomó en consideración la solidez del modelo para la eliminación de los compuestos (robustez LSO). Este aspecto fue considerado como la variación del porcentaje de buena clasificación del modelo modificado (Δ) con el LSO con respecto al original. El promedio de previsibilidad LSO fue del $86,41 \pm 0,95\%$ (promedio \pm SD) y $\Delta = -0,55\%$, siendo 6 el número promedio de medicamentos analizados contra cada especie de hongos.

Los resultados de algunas de las 87 especies estudiadas fueron *Candida albicans*: 43 compuestos analizados, el 100% de la previsibilidad LSO, $\Delta = -3,49\%$; *Candida parapsilosis*: 23, 100%, $\Delta = -0,86\%$; *Aspergillus fumigatus* 21, 95,20%, $\Delta = 0,05\%$; *Microsporium canis* 12,

91,60%, $\Delta = -2,84\%$; *Trichophyton mentagrophytes* 11, 100%, $\Delta = -0,51\%$; *Cryptococcus neoformans* 10, 90%, $\Delta = -0,90\%$. La ecuación del modelo es la siguiente:

$$\begin{aligned} Actv = & -2.88 \cdot {}^0C_s(X) + 1.26 \cdot {}^5C_s(X) - 1.01 \cdot {}^0C_s(T) \\ & - 0.78 \cdot {}^0C_s(C_{uns}) + 0.94 \cdot {}^3C_s(X) - 0.76 \cdot {}^4C_s(T) \\ & - 1.17 \\ & \lambda = 0.53; \quad F(6,484) = 71.93; \quad p < 0.001 \end{aligned} \quad (8)$$

donde λ es la estadística de Wilk, la estadística de la discriminación total, F es la relación de Fisher, y p es el nivel de error. En esta ecuación, kC_s se calcula para la totalidad (T) de átomos en la molécula o para asociaciones específicas de átomos. Estas asociaciones son átomos con una característica común: X: halógenos y C_{uns} : átomos de carbono insaturados.

González-Díaz y Prado-Prado [80] han seleccionado pares de medicamentos antifúngicos con perfil de similares/diferentes especies para predecir la actividad y las representaron como una gran red. A continuación, desarrollaron un modelo de clasificación mt-QSAR, en el que los resultados fueron las entradas de esta red. La precisión general de la clasificación del modelo fue del 87,0% (161 de los 185 compuestos) en entrenamiento, del 83,4% (50 de los 61) en validación, y del 83,7% para 288 compuestos antifúngicos adicionales utilizados para extender la validación del modelo para la construcción de la red. La red prevista tiene 59 nodos (compuestos), 648 aristas (pares de compuestos con actividad similar), baja densidad de cobertura $d = 37,8\%$, y una distribución más cercana a un valor normal que a uno exponencial. La ecuación del modelo es la siguiente:

$$\begin{aligned} Actv = & -0.49 \cdot {}^A\pi_5(s, C_{sp\&sp2}) - 2.57 \cdot {}^A\pi_0(s, X) + 1.43 \cdot {}^A\pi_0(s, H - Het) + 0.90 \\ & R_c = 0.75 \quad \lambda = 0,44 \quad p < 0.001 \end{aligned} \quad (9)$$

donde R_c es el coeficiente de correlación canónica, λ es la estadística de Wilk, y p el nivel de error. En esta ecuación, las probabilidades absolutas ${}^A\pi_k$ calculadas se refieren a:

1. ${}^A\pi_5(s, C_{sp\&sp2})$ todos los átomos de carbono insaturados (átomos sp y sp2) y todos los átomos colocados a una distancia de cinco o menos átomos de ellos.
2. ${}^A\pi_0(s, X)$ todos los átomos de halógenos.
3. ${}^A\pi_0(s, H-Het)$ todos los átomos de hidrógeno unidos a un heteroátomo (N, O, o S).

Prado-Prado *et al.* [81] han utilizado la teoría de las Cadenas de Markov para calcular nuevos momentos espectrales para múltiples dianas con el fin de ajustar un modelo mt-QSAR

que predice la actividad antifúngica de más de 280 medicamentos contra 90 especies de hongos. El LDA se utilizó para clasificar los medicamentos como activos o inactivos contra especies de hongos diferentes. El modelo clasificó correctamente 12434 de los 12566 compuestos inactivos (98,95%) y 421 de los 468 compuestos activos (89,96%). La predictibilidad total para el entrenamiento fue del 98,63%. La validación del modelo se llevó a cabo mediante series de predicción externas, clasificando 6216 de los 6277 compuestos inactivos y 215 de los 239 compuestos activos. La predictibilidad total en el entrenamiento fue del 98,7%. La ecuación del modelo es la siguiente:

$$\begin{aligned}
 Actv = & -3.44 \cdot^5 \mu_s(\text{Het}) - 3.18 \cdot^2 \mu_s(\text{H - Het}) - 3.85 \cdot^3 \mu_s(\text{C}_{\text{sat}}) \\
 & + 4.76 \cdot^4 \mu_s(\text{C}_{\text{sat}}) - 4.61 \cdot^5 \mu_s(\text{C}_{\text{sat}}) + 28.26 \cdot^0 \mu_s(\text{T}) - 29.26 \\
 & \lambda = 0.33; \quad ^2\chi = 14367.94; \quad p < 0.001
 \end{aligned}
 \tag{10}$$

donde, χ^2 es el Chi-cuadrado, y p el nivel de error. En esta ecuación, $^k\mu_s$ se calcularon para el total (T) de átomos en la molécula o para asociaciones específicas de átomos. Estas asociaciones son átomos con una característica común: Het: heteroátomo, H-Het: hidrógeno unido a heteroátomos, C_{sat} : átomos de carbono saturados.

1.3. Herramientas online de clasificación molecular

En la sección anterior hemos presentado modelos QSAR para compuestos anti-virales, anti-bacterianos, anti-parásitos y anti-fúngicos. Estos modelos no están implementados en servidores Web como la mayoría de los modelos QSAR en la literatura. En la sección actual presentamos algunos ejemplos de páginas Web con modelos tipo QSAR con aplicaciones en Microbiología y Parasitología.

La localización de las proteínas en virus y bacterias es muy importante para el desarrollo de fármacos nuevos y en la búsqueda de dianas moleculares. Por ello, el grupo de Kuo-Chen Chou (http://www.csbio.sjtu.edu.cn/index_eng.htm) propone tres servidores online para la predicción de la ubicación de las proteínas en los virus, bacterias gram-negativas y gram-positivas.

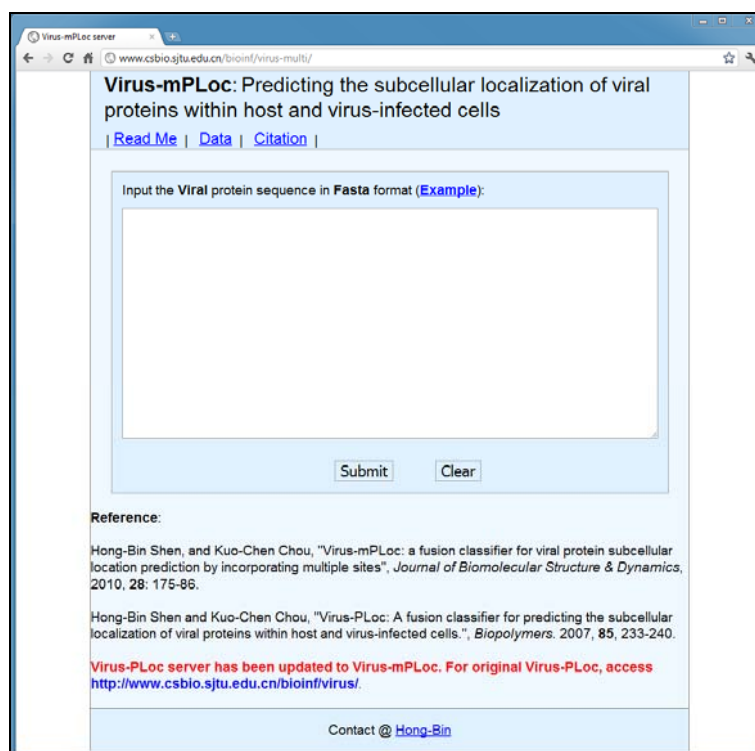


Figura 9: El servidor *Virus-mPLoc* para predecir la ubicación subcelular de las proteínas virales

El primer servidor es *Virus-mPLoc* [82] y sirve para predecir la ubicación subcelular de las proteínas virales utilizando la información de diversos sitios Web (**Figura 9**). El conocimiento de la ubicación subcelular de las proteínas virales en una célula huésped o en las células infectadas por un virus es muy importante porque está relacionado con sus tendencias

destructivas y sus consecuencias. Frente a la avalancha de nuevas secuencias de la proteína descubierta en la era post-genómica, nos enfrentamos al reto de desarrollar métodos automatizados de forma rápida y precisa para la predicción de los sitios de ubicación de las proteínas virales en una célula huésped; la información adquirida es particularmente importante para la ciencia médica y el diseño de fármacos antivirales. Shen *et al.* desarrollaron un clasificador de fusión llamado *Virus-mPLOC* establecido por la hibridación de la información genética de Ontología, la información del dominio funcional y la información de la evolución secuencial. La nueva herramienta no sólo puede predecir con mayor exactitud los sitios de ubicación de las proteínas virales en una célula huésped, sino que también tiene la capacidad de identificar la ubicación de varias proteínas del virus, que está más allá del alcance de cualquier predicción existente especializada en proteínas virales. El servidor está implementado en <http://www.csbio.sjtu.edu.cn/bioinf/virus-multi/>.

El segundo servidor, *Gneg-mPLOC* [83], predice la ubicación de las proteínas en bacterias gram-negativas incorporando la información de ontología de los genes, el dominio funcional y la evolución secuencial (**Figura 10**). Se puede utilizar para identificar proteínas en bacterias Gram-negativas en ocho ubicaciones: (1) citoplasma, (2) extracelular, (3) *fimbrium*, (4) flagelo, (5) membrana interna, (6) nucleído, (7) membrana externa, y (8) periplasma. También se puede utilizar para el caso en que una proteína de una búsqueda puede existir al mismo tiempo en más de un lugar. En comparación con el pronóstico original llamado *Gneg-Ploc*, el nuevo modelo es mucho más potente y flexible. Para un conjunto de datos de referencia en los que ninguna de las proteínas ha incluido una identidad de secuencia más del 25% en comparación con otras de la misma ubicación, la clasificación *Gneg-mPLOC* fue del 85,5%, que era más de un 14% superior a la tasa correspondiente al *Gneg-Ploc*. Como servidor gratuito, *Gneg-mPLOC* se encuentra en <http://www.csbio.sjtu.edu.cn/bioinf/Gneg-multi/>.



Figura 10: El servidor *Gneg-mPLOC* para predecir la ubicación de las proteínas en bacterias gram-negativas

El tercer servidor *Gpos-mPLOC* [84] es similar al *Gneg-mPLOC*, sirve para predecir la ubicación de las proteínas en bacterias gram-positivas y está implementado en <http://www.csbio.sjtu.edu.cn/bioinf/Gpos-multi/>.

Otro ejemplo de servidor para los virus es *HIVcleave* [85], una herramienta para predecir los sitios de *cleavage* de las proteasas del HIV (virus de inmunodeficiencia humana) en proteínas. Según la "teoría de la clave distorsionada" [86], la información de los sitios de escisión (*cleavage*) de las proteínas por la proteasa del HIV es muy útil para encontrar inhibidores eficaces contra el HIV, la causa del SIDA (síndrome de inmunodeficiencia adquirida). Para satisfacer la creciente necesidad en este sentido, se ha implementado este servidor web en <http://chou.med.harvard.edu/bioinf/HIV/> (**Figura 11**). Se ofrece también una

guía online paso-a-paso sobre cómo utilizar *HIVcleave* para identificar los sitios de corte para una consulta de secuencias de proteínas por las proteasas del HIV-1 y del HIV-2.

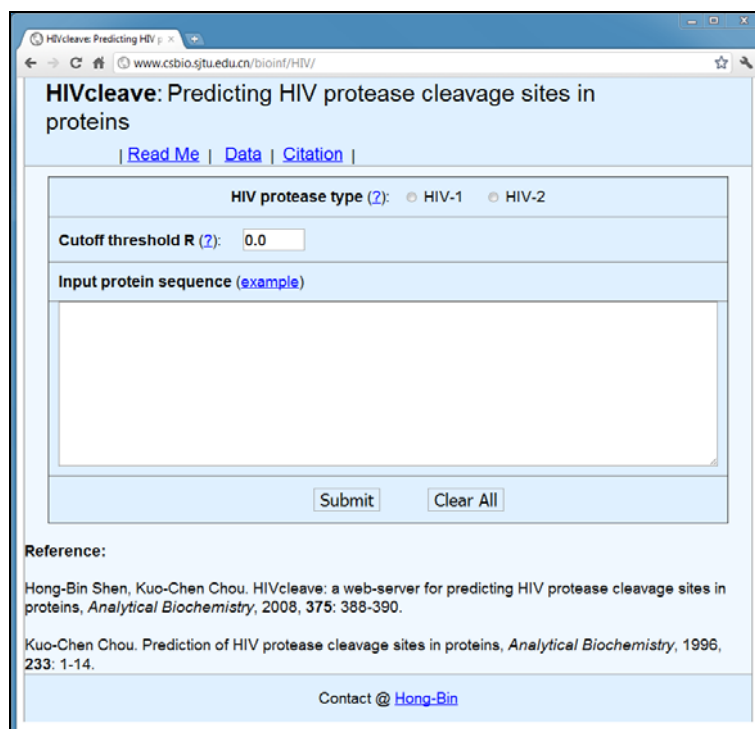


Figura 11: El servidor *HIVcleave* para predecir los sitios de *cleavage* de las proteasas del HIV en proteínas

Una colección de modelos QSAR para diversos organismos como dianas está presentada en la Web del *Open QSAR* (<http://www.openqsar.org>). Aquí se pueden encontrar ejemplos de modelos validados y estables con técnicas lineales, redes neurales artificiales (ANN) y de regresión mediante mínimos cuadrados parciales (PLS) para organismos como los virus (Human herpesvirus, Hepatitis C virus, HIV-1), *Entamoeba histolytica*, *Leishmania donovani*, *Plasmodium falciparum* y *Toxoplasma gondii*. Las desventajas de estos modelos son el número muy reducido de casos usados para entrenar y para validar el modelo.

El número reducido de herramientas online con modelos QSAR para el descubrimiento de fármacos y sus dianas proteicas correspondientes ha creado la necesidad de nuevos servidores públicos. En esta tesis se presentan en la sección “**2.2. Nuevos servidores online del Bio-AIMS** basados en técnicas de ingeniería informática e inteligencia artificial” una colección de 7 implementaciones de modelos QSAR para fármacos y proteínas con aplicaciones para al menos 9 tipos de microbios y parásitos: *Ascaris*, *Entamoeba*, *Fasciola*, *Giardia*, *Leishmania*, *Plasmodium*, *Trichomonas*, *Tripanosoma* y *Toxoplasma*.

1.4. Objetivos

- ❖ **Diseño aplicaciones**: desarrollar nuevas herramientas informáticas (programas de ordenador) con técnicas de ingeniería informática para el cálculo de TIs, de utilidad en el desarrollo de modelos QSAR.
- ❖ **Modelos QSAR/QSPR**: encontrar nuevos modelos QSAR/QSPR con técnicas de inteligencia artificial aplicables a la predicción de la actividad biológica de compuestos de interés en Química Farmacéutica, Microbiología y Parasitología empleando los nuevos programas desarrollados.
- ❖ **Diseño servidores online**: implementar los cuatro modelos QSAR/QSPR encontrados en nuevas herramientas informáticas de uso en la red (servidores web) para la predicción online de fármacos y dianas moleculares en Química Farmacéutica, Microbiología y Parasitología.
- ❖ **Publicaciones**: protección de la propiedad intelectual (registros de software), comunicación (publicación de artículos, libros, capítulos, etc.) y aplicación de las herramientas desarrolladas.

2. RESULTADOS Y DISCUSIÓN

En esta sección se presentarán todos los resultados obtenidos en forma de artículos de revisión y capítulos de libro, manuales para las herramientas informáticas desarrolladas (programas de ordenador) y artículos de investigación ya publicados por el autor.

Los tres programas de ordenador desarrollados y/o registrados fueron: MInD-Prot, S2SNet y CULSPIN. Se presenta un total de 6 publicaciones científicas (artículos de revista con índice de impacto JCR), agrupadas según el objetivo específico que cumplimentan. Los cuatro servidores web (herramientas de uso online) desarrollados fueron: Trypano-PPI, Plasmod-PPI, ATCUNpred y LIBPpred. Los servidores finales fueron utilizados con el fin de apoyar los datos experimentales para más de nueve tipos de parásitos como son *Trypanosoma*, *Plasmodium*, *Trypanosoma*, *Leishmania*, *Toxoplasma*, *Shigella* y *Cryptosporidium*. A cada servidor web le corresponde un artículo publicado en el que se describe el desarrollo, la validación, y la aplicación de la herramienta. En otros artículos se describen las metodologías y/o los algoritmos que fueron necesarios desarrollar previamente para la creación de los servidores presentados. Para cada artículo se presenta una breve sección explicativa en español de su importancia y los resultados alcanzados. En el apartado “**5. PUBLICACIONES (ANEXOS)**” de esta Tesis se adjuntan las publicaciones correspondientes en el idioma en el que fueron publicadas.

2.1. Nuevos programas de ordenador para los parámetros moleculares

Para predecir las actividades biológicas de los fármacos o para buscar las dianas moleculares con modelos QSAR/QSPR se necesitan números con el fin de caracterizar cuantificamente la relación entre la estructura de las moléculas y las actividades biológicas. Por ello, se han desarrollado nuevos programas para ordenador capaces de calcular descriptores moleculares/índices topológicos para fármacos, proteínas, ácidos nucleicos u otros sistemas reales: MInD-Prot (similar a las funciones de MARCH-INSIDE), S2SNet (para grafos de tipo estrella) y CULSPIN (para grafos de tipo espiral). Las funciones MInD-Prot han sido utilizadas en la implementación online de las herramientas presentadas en “2.2. Nuevos servidores online del Bio-AIMS basados en técnicas de ingeniería informática e inteligencia artificial” y, en consecuencia, las publicaciones donde se ha utilizado este programa están presentadas con cada servidor. Los otros dos programas, S2SNet y CULSPIN se presentan en tres partes: las publicaciones con la aplicación, el manual del programa y el certificado de registro general de la propiedad intelectual.

2.1.1. MInD-Prot – Descriptores Markov para fármacos y proteínas

MInD-Prot (Markov Inside for Drugs and Proteins = índices tipo Markov para fármacos y proteínas) es una aplicación programada en Python/wxPython para el cálculo de los siguientes índices tipo Markov para fármacos y proteínas:

- Momentos espectrales y entropías Shannon (sólo para las proteínas)
- Propiedad promedio (para los fármacos y las proteínas)

La aplicación (**Figure 12**) puede calcular los índices promedios para las redes complejas de las moléculas de proteínas y fármacos, mediante el uso de las clases de entrada (para los medicamentos y proteínas) o de la información de los PDBs (sólo para las proteínas). Además, MInD-Prot puede generar los índices mezclados de pares de proteínas, pares de fármacos o pares proteína-fármaco. Si es necesario, la herramienta puede generar al azar pares negativos para los pares de proteína y proteína-fármaco. Se puede obtener información adicional, como son las cabeceras (*headings*) de los PDBs para las proteínas. Estos números que caracterizan a cada proteína/fármaco o a un par proteína-fármaco se utilizan para la construcción de modelos de clasificación tipo QSAR/QPDR.

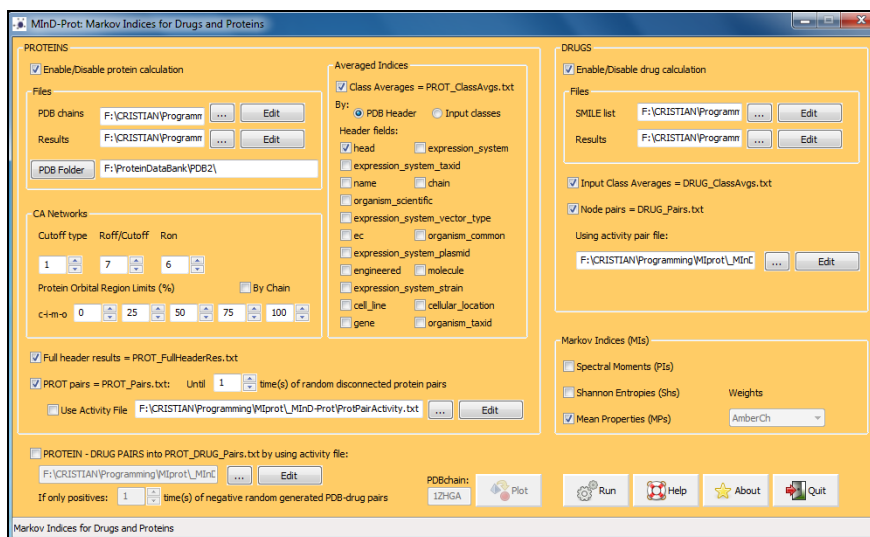


Figura 12: Interfaz programa MInD-Prot

Como utilizar el MInD-Prot

En la ventana principal se pueden elegir los parámetros de cálculo y los parámetros de entrada/salida.

La interfaz principal del usuario se divide en las siguientes partes:

Proteínas (*PROTEINS*):

- Archivos de parámetros
 - Archivo de entrada con la lista de cadenas de proteínas o el nombre de las proteínas de la base de datos PDB Databank (<http://www.pdb.org>)
 - Archivo con los resultados simples para el cálculo de las proteínas
 - Carpeta local con los archivos PDB para las proteínas de entrada; si no existe el PDB, se descargará automáticamente desde la Web
- Parámetros para la red de carbonos alpha de los aminoácidos
 - Los parámetros Cutoff, Roff, Ron para definir las condiciones para considerar unidos dos átomos de carbono alpha de los aminoácidos (definición de la red compleja para cada proteína)
 - Límites para los orbitales proteicos en %: core, inner, middle, outer
 - Atributo “By Chain”: si está activo, el cálculo buscará todas las cadenas para las proteínas; si está inactivo, el cálculo considerará la proteína entera
- Índices promedios (*Averaged Indices*) = PROT_ClassAves.txt
 - Utilizando los campos del *header* del PDBs: head, expression_system, expression_system_taxid, name, chain, organism_scientific, molecule, expression_system_vector_type, ec, organism_common,

expression_system_plasmid, engineered, expression_system_strain, cell_line,
cellular_location, gen, organism_taxid

- Utilizando las clases desde el archivo de entrada (*Input classes*)
- Salida con el header entero (*Full header information output*) =
PROT_FullHeaderRes.txt
 - obtener la información completa del *header* de PDB y añadirla al resultado simple (una columna para cada campo de la cabecera)
- Pares de proteínas (*Protein PAIRS*)
 - Utilizando la similitud de las cadenas de PDB como pares positivos y generando casos negativos hasta X veces los pares positivos
 - Como alternativa, se puede utilizar un archivo con actividades (*activity file*) (predefinido: ProtPairActivity.txt) con PDB1[tab]PDB2[tab]Class

Fármacos (*DRUGS*)

- Parámetros archivos (*Files*)
 - Archivo con los códigos SMILES como Drug Name[tab]SMILE formula
 - Archivo para la salida simple con los cálculos de los índices topológicos para los fármacos
- Resultados promedios utilizando las clases de los archivos de entrada (*Averaged results by input classes*) = DRUG_ClassAvg.txt con Drug Name[tab]SMILE formula[tab]Class
- Pares de fármacos (*Drug PAIRS*): siempre se hacen utilizando un archivo con la actividad biológica de los fármacos como DrugName1[tab]DrugName2[tab]Class

Índices tipo Markov (*Markov Indices*)

- ❖ Existen tres tipos de índices: momentos espectrales (*Spectral Moments*), entropías tipo Shannon (*Shannon Entropies*) y propiedades promedias (*Mean Properties*)
- ❖ Se puede calcular separado las proteínas y los fármacos; si se calculan los dos, se utiliza automáticamente sólo Mean Properties basados en la electronegatividad tipo Amber de los átomos (AmberCh). Para el cálculo de las proteínas, se pueden utilizar otras propiedades de los átomos/aminoácidos, tales como Polar_KJ, AtContrib2P, AtRefr, vdWArea, hardness_I-A, Electrophilicity, ElectroMulliken y los otros tipos de índices: Spectral Moments and Shannon Entropies.

Pares proteína-fármaco (*PROTEIN-DRUG PAIRS*)

- Utilizando un archivo de entrada con la propiedad del par proteína- fármaco con PDBChain[tab]DrugName[tab]Activity
 - Si existe sólo un tipo de actividad (casos positivos), se pueden generar al azar pares de proteína-fármaco hasta X veces los casos positivos
- Se puede calcular este tipo de pares sólo si los dos tipos de cálculos para las proteínas y para los fármacos están activados

Notas: Los archivos de entrada y los de salida se pueden crear/modificar directamente en la interfaz utilizando el NotePad nativo de Windows.

Los índices Markov del MInD-Prot

Antes de calcular los índices, la matriz con las conectividades del grafo molecular estará normalizada (tipo Markov = los elementos de la matriz se dividen con el máximo valor de su fila), resultando una matriz con las probabilidades de los nodos (**P**). En un segundo paso, P será elevado al poder ($k=5$ veces), resultando k matrices (**P k**), la entrada para el cálculo de los índices:

- **Spectral Moments** (PI);
- **Shannon Entropy** (Sh);
- **Mean Properties** (MP).

El MInD-Prot calcula los índices de modo similar al MARCH-INSIDE pero sin conseguir el efecto del entorno molecular. Las ventajas del MInD-Prot son las siguientes:

- ✓ para proteínas:
 - cálculo de índices promedios para cada tipo de clase de proteína de la entrada o utilizando los campos del *heading* de los archivos PDB;
 - extraer toda la información de las cabeceras de los PDBs para cada proteína, al lado de los índices;
 - crear interacciones de proteína-proteína utilizando para las cadenas de la misma proteína y generar pares negativos al azar;
 - crear interacciones de proteína-proteína utilizando las clases de proteínas de la entrada y generar pares negativos al azar;
 - cálculo de índices mixtos para pares de proteínas;

- ✓ para fármacos:
 - cálculo de índices promedios para cada tipo de clase de fármaco de la entrada;
 - crear pares de fármaco – fármaco utilizando los actividades biológicas de la entrada;
 - cálculo de índices mixtos para pares de fármacos;
- ✓ para proteínas y fármacos:
 - crear pares de proteína – fármaco utilizando las interacciones entre ellos y generar pares negativos al azar;
 - cálculo de índices mixtos proteína – fármaco promedios por orbita de la proteína y por índice k para los fármacos;
- ✓ los pares proteína – proteína, fármaco – fármaco y proteína – fármaco forman redes complejas de interacciones muy útiles en el descubrimiento de nuevos fármacos y sus dianas moleculares correspondientes.

2.1.2. S2SNet – Índices topológicos del grafo de tipo estrella

2.1.2.1. Publicaciones con S2SNet

2.1.2.1.1. Clasificación tipo Random Forest basada en los índices topológicos del grafo tipo estrella de las proteínas antioxidantes

Random Forest Classification based on Star Graph Topological Indices for Antioxidant Proteins
[Journal of Theoretical Biology](#) 317, 331-337 (2013)

Enrique Fernandez-Blanco, Vanessa Aguiar-Pulido, **Cristian R Munteanu**, Julian Dorado

Enlace: <http://goo.gl/R5vV8>

Envejecimiento y calidad de vida es un tema de investigación importante hoy en día en áreas como las ciencias biológicas, química, farmacología, etc. La gente vive más tiempo y quiere pasar ese tiempo con una mejor calidad de vida. En este sentido, existe un pequeño subconjunto de moléculas en la naturaleza, llamado proteínas antioxidantes, que pueden influir en el proceso de envejecimiento. Sin embargo, la prueba de cada proteína individual con el fin de identificar sus propiedades es bastante cara e ineficiente. Por esta razón, este trabajo propone un modelo en el que la estructura primaria de la proteína se representa mediante los gráficos de redes complejas, que se pueden utilizar para reducir el número de proteínas sometidas a ensayo para establecer su actividad biológica antioxidante. El gráfico obtenido como una representación teórica de una proteína ayudará a describir el sistema complejo mediante el uso de índices topológicos. Más específicamente, en este trabajo, se han sido utilizado redes tipo estrella, así como los índices correspondientes, calculados con la herramienta S2SNet. Con el fin de simular la proporción existente de proteínas antioxidantes en la naturaleza, se ha creado un conjunto de datos que contiene 1999 proteínas, de las cuales 324 son proteínas antioxidantes. Con estos datos como entrada, los índices topológicos de los gráficos estrella se calcularon con la herramienta S2SNet. Estos índices, se utilizan luego como entrada en varias técnicas de clasificación. Entre las técnicas utilizadas, el Random Forest ha mostrado el mejor rendimiento, logrando una puntuación de 94% de casos totales correctamente clasificados. El modelo propuesto es capaz de alcanzar un porcentaje de 81,8% de casos clasificados correctamente para el grupo de las proteínas antioxidantes, con una precisión del 81,3%.

2.1.2.2. Manual del programa S2SNet

Lenguaje de S2SNet

La S2SNet (*Sequence to Star Network*) es una aplicación gratuita en el campo de las redes complejas (matemáticas aplicadas) programada en el lenguaje Python, utilizando el wxPython para crear el entorno gráfico y los ejecutables del *Graphviz* para dibujar los grafos (<http://www.graphviz.org/>). La ayuda está presentada como una página de HTML. La S2SNet funciona en el sistema operativo Microsoft XP/Vista. Para editar los archivos de cálculos se utiliza el editor Bloc de Notas.

Nota: en los dos casos se necesita la instalación previa del *Graphviz* para la visualización de los grafos.

La S2SNet – aplicación para estudios de redes complejas:

- ✓ **lenguaje de programación:** Python, wxPython, HTML;
- ✓ **sistema operativo:** Microsoft XP y Vista.
- ✓ **aplicaciones externas:**
 - ejecutables de Graphviz: *dot*, *circo*, *twopi*, *neato* y *fdp*;
 - *Bloc de Notas* de MS Windows XP/Vista (*Notepad*).

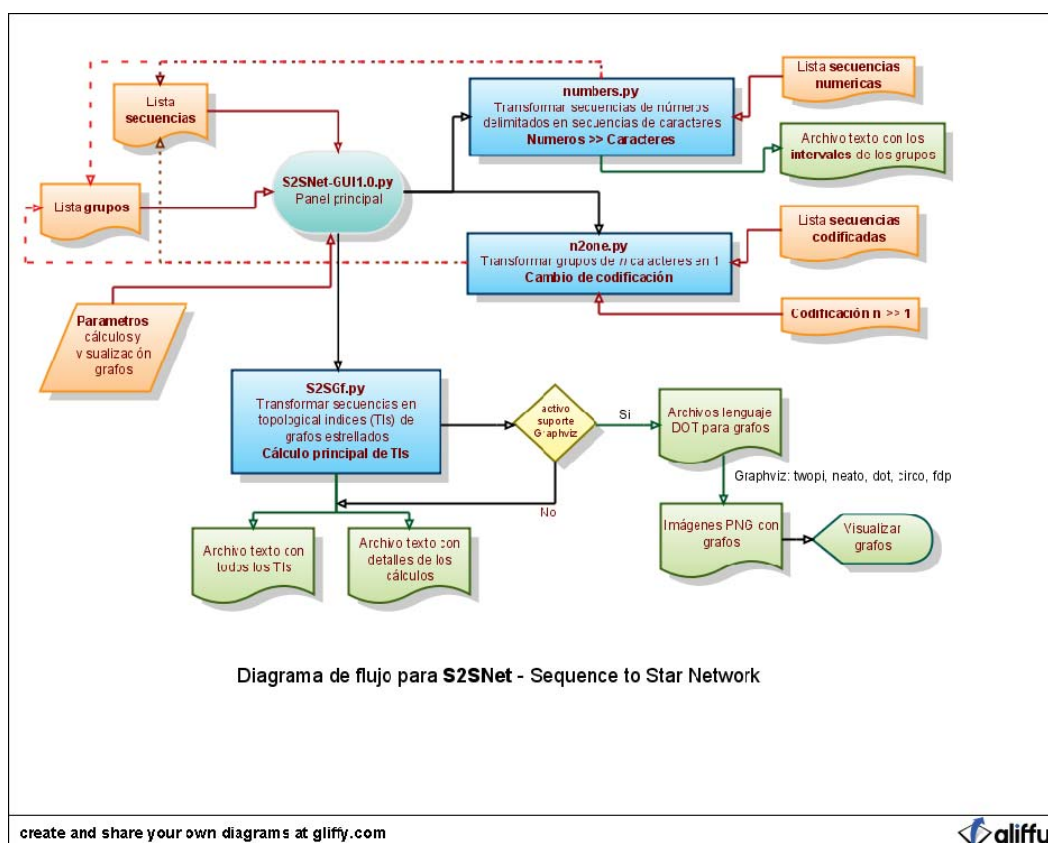


Figura 13: Diagrama lógica de la S2SNet

¿Qué es la S2SNet?

La S2SNet transforma secuencias de caracteres en índices topológicos (TIs) de redes complejas de tipo estrella (*Star Network*, SN) y visualiza los grafos resultados (**Figura 13**). Con estos índices se pueden realizar diversos análisis estadísticos o crear modelos QSAR (relación estructura - propiedades). Ejemplos de secuencias son las cadenas de aminoácidos de las proteínas, los ácidos nucleicos y los espectros de masa de proteínas. La S2SNet se puede utilizar para estudiar distintos sistemas, desde sistemas simples de átomos en pequeñas moléculas anti-cancerígenas, hasta sistemas complejos de redes metabólicas, sociales, computacionales o sistemas biológicos.

¿Qué puede hacer la S2SNet?

- ✓ **Transformar** las secuencias en **índices topológicos de redes de tipo estrella** (menú Calculations, *Sequence to Star Network* o el botón S2SNet desde el panel principal del programa): las **entropías de Shannon** de los n matrices Markov (Sh), **traces** de las mismas matrices (Tr), el número de **Harary** (H), el índice de **Wiener** (W), los índices topológicos de **Gutman** ($S6$), de **Schultz** (non-trivial part) (S), de

Moreau-Broto (ATS_n), el índice de conectividad de distancia **Balaban** (J), los índices de conectividad **Kier-Hall** y **Randic**;

- ✓ **Transformar** los datos de tipo numérico en secuencias de caracteres (menú Calculations, *Numbers to Sequence*);
- ✓ **Transformar** las secuencias de grupos de n caracteres en secuencias simples como un cambio de codificación (menú Calculations, *N to 1-Character Sequence*);
- ✓ **Editar/Visualizar** los archivos de entrada y de salida de tipo texto;
- ✓ **Crear** archivos que describen grafos en el lenguaje **DOT**; estos archivos se utilizan como entrada de los ejecutables de Graphviz para visualizar los grafos;
- ✓ Crear imágenes PNG con los grafos y visualizarlas.

Descripción de la S2SNet

La S2SNet es un programa interactivo que tiene dos paneles: el panel principal y la consola de DOS (**Figura 14**).

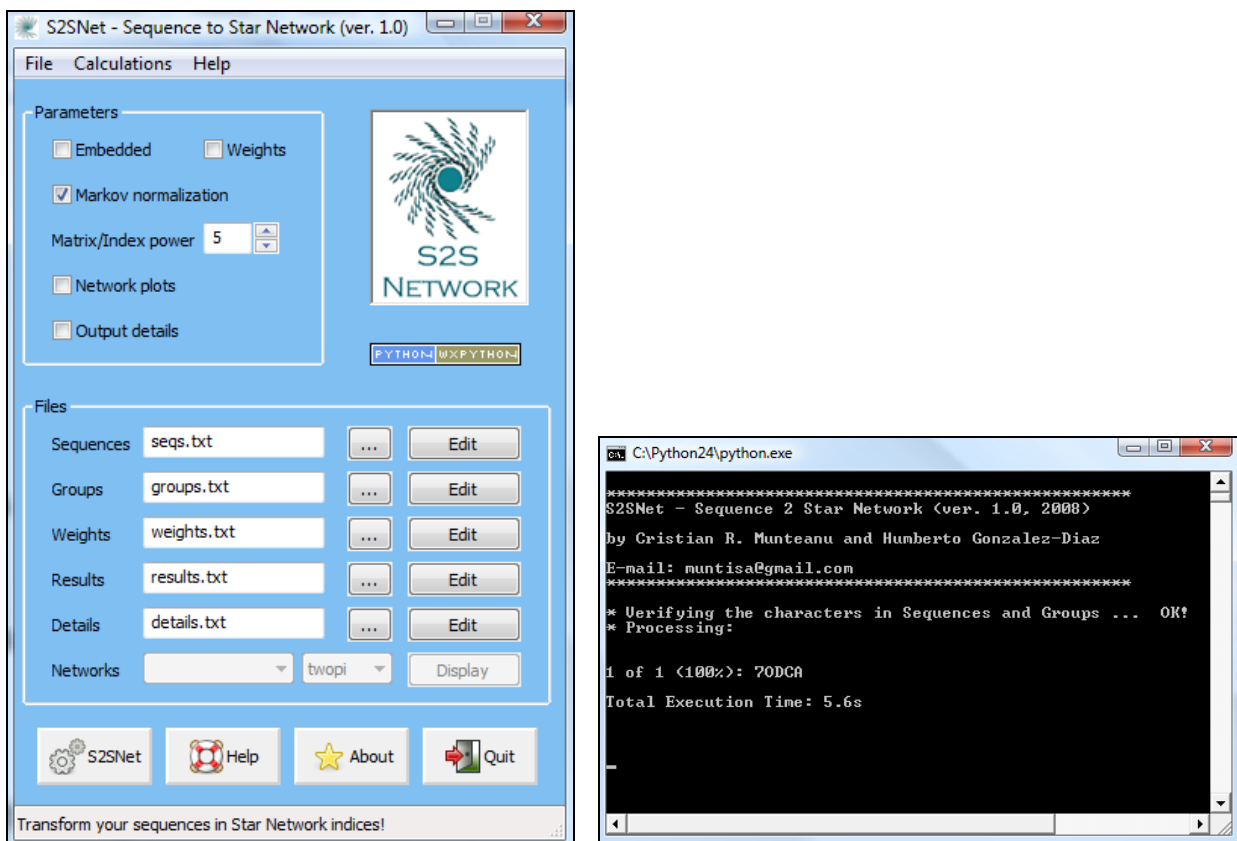


Figura 14: Interfaz de la S2SNet

Además, la S2SNet tiene una ayuda (*Help*), detalles sobre el programa y los autores (*About*), la posibilidad de crear un archivo de texto nuevo (*New*) y la opción de salir de la aplicación (*Quit*). Los botones se doblan con las opciones en los menús.

En la consola DOS se pueden ver siempre el estado de los cálculos y los errores.

¿Cómo se utiliza la S2SNet?

En la ventana principal se pueden elegir los parámetros del cálculo de los índices topológicos específicos, los archivos de entrada/salida y el tipo de visualización de los grafos:

- ✓ **Parámetros:** *embedded* se utiliza para crear redes/grafos *embebidos*; *weight* se utiliza para poner valores de peso en los nodos de los grafos; la normalización de tipo *Markov* para las matrices de conectividad; si se quieren ver los detalles del cálculo se elige *details*; el poder de las matrices de conectividad con el valor de *power* (máx. 5); el soporte para la creación y visualización de los grafos con la opción *Network plots*;
- ✓ **Archivos de entrada:** *sequences*=secuencias, *groups*=grupos y *weights*=pesos;
- ✓ **Archivos de salidas / resultados:** *results*=resultados y *details*=detalles;
- ✓ **Display mode** = el tipo de visualizar la red: *sequence*=el nombre de la secuencia y el tipo del ejecutable de Graphviz (**dot**, **circo**, **twopi**, **neato** y **fdp**); se calculan automáticamente los grafos máximos y promedios de todas las secuencias analizadas.

Ejemplo de cálculo con la S2SNet

Un ejemplo de cálculo es utilizar una secuencia proteica, 7ODCA, de la base de datos de proteínas, Protein Data Bank (<http://www.rcsb.org/>). Las entradas con la secuencia de aminoácidos y los grupos se presentan en la **Figura 15**.

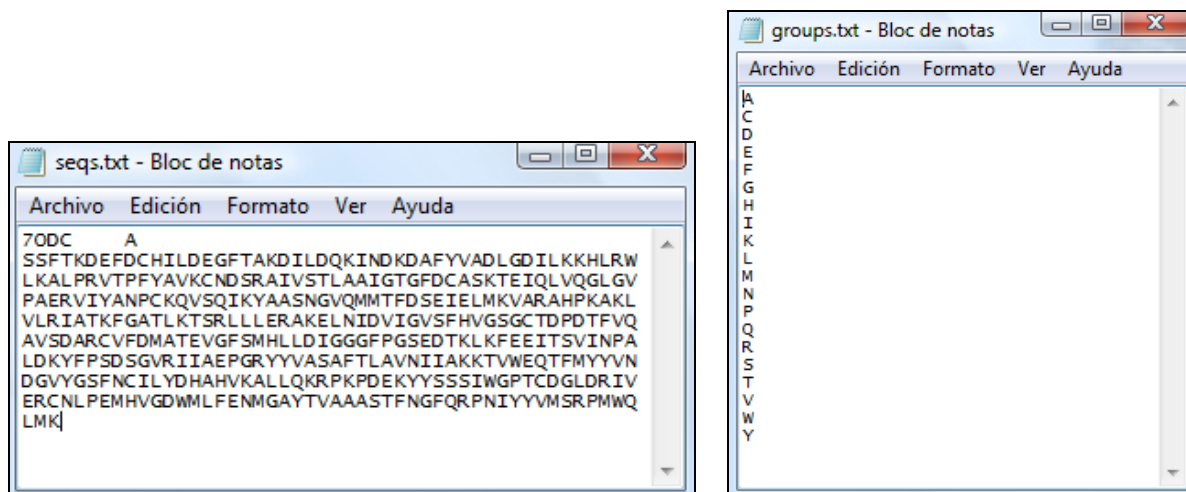


Figura 15: Ejemplo de entrada en la S2SNet (secuencia y grupos de aminoacidos)

La S2SNet transforma la secuencia en una lista de índices topológicos específicos para el grafo de tipo estrella y también puede generar las imágenes de los grafos con la ayuda del Graphviz.

En la **Figura 16** se presentan los resultados para los cálculos de grafos “non-embedded” (grafo situado a la izquierda con *neato*, grafo situado a la derecha con *dot*). En la **Figura 17** se presenta el caso de los grafos “embedded”, los resultados se modifican incluyendo en los cálculos la conectividad inicial dentro de la secuencia (grafo situado a la izquierda con *twopi*, grafo situado a la derecha con *circo*).

The image shows a Notepad window titled 'results.txt - Bloc de notas' containing the following data:

PDB	Chain	Seq	Sh0	Sh1	Sh2	Sh3	Sh4	Sh5
Tr0	Tr1	Tr2	Tr3	Tr4	Tr5	H	W	S6
S	J	X0	X1(R)	X2	X3	X4	X5	
70DC	A							
SSFTKDEFDC HILDEGFTAKDILDQKINDKDAFYVADLGDILKKHLRWLKALPRVTPFYAVKCNDSRAIVSTLAAIGTGFDCAASKTEIQLVQGLGVPAERVIYANPCQVVSQIKYAAASNGVQMMTFDSEIELMKVARAHPKAKLVLR IATKFGATLKTSRLLERAKELNIDVIGVSFHVSGGCTDPDTFVQAVSDARCVFDMATEVGFMSHLLDIGGGFPGSEDTKLKFEIITSVINPALDKYFPDSGVR IIAEPGRYYVASAFTLAVNIIAKKTVEQTFMYVYNDGVYGSFNCILYDHAHVKALLQKRPKPDEKYYSSSIWGPTCDGLDR.IVERCNLPEMHVGDWMLFENMGAYTVAASTFNGFQRPNIIYYVMSRPMWQLMK								
5.91683715007	6.80911581588	6.41546759481	6.87788629113	6.55167891758				
6.90893894584	388	0.0	194.5	0.0	146.125	0.0		
1569.1127847	9926318.0		13322.3405011	39747914.0				
7680365149.45	279.731795493		190.804413284	127.848026702				
85.4022066419	56.8529455389		37.6656998544					

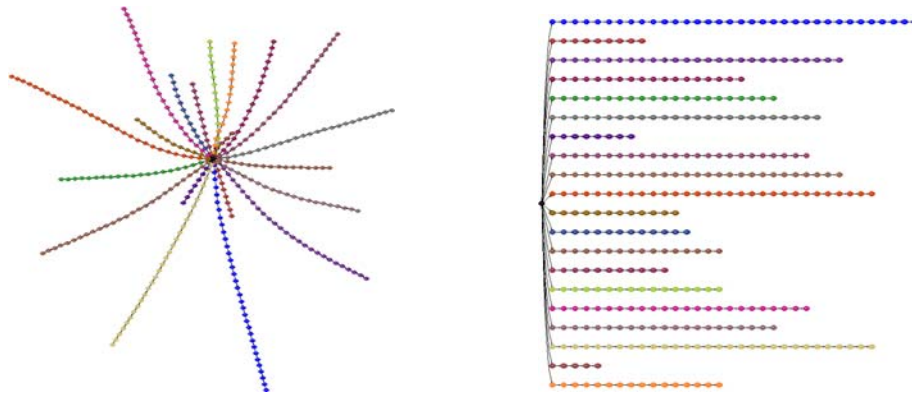


Figura 16: Ejemplo de resultados “non-embedded” con la S2SNet: índices topológicos y dibujos de los grafos de tipo estrella

Archivo	Edición	Formato	Ver	Ayuda					
PDB	Chain	Seq	Sh0	Sh1	Sh2	Sh3	Sh4	Sh5	
Tr0	Tr1	Tr2	Tr3	Tr4	Tr5	H	W	S6	
S	J	X0	X1(R)	X2	X3	X4	X5		
70DC	A								
SSFTKDEFDCCHILDEGFTA KDI LDQKINDKDAFYVADLGDILKKHLRWL KALPRVTFYAVKCNDSRA IVSTLAAIGTGFDCASKTEIQLVQGLGVAERVIYANPCKQVSIKYAASNGVQMMTFDSEIELMKVA RAHPKAKLVLR IATKFGATL KTSRLLERAKELNIDVIGVSHVSGCTDPDFTFVQAVSDARCVFDMA TEVGF SMHL LDIGGGFPGSEDTKLKFEIITSVINPALDKYFPSDSGVR.IIAEPGRYYVASAFTLAVNI IAKKTVWEQTFMYVNDGVYGSFNCILYDHAHVKALLQKRPKPDEKYYYSSSIWGP TCDGLDRIVERCN LPEMHVGDWMLFENMGAYTVA AASTFNGFQRPNIIYYVMSRPMWQLMK 5.94277935927 6.20302610201 6.15115323348 6.20767598017 6.18925683378 6.21090446699 388 0.0 101.033333333 1.48125 46.8350462963 2.29915123457 2149.32953498 9735114.0 64058.5120342 75274220.0 77495217.4044 199.402675045 192.710325405 184.618394946 176.731552801 168.771638841 160.909639012									

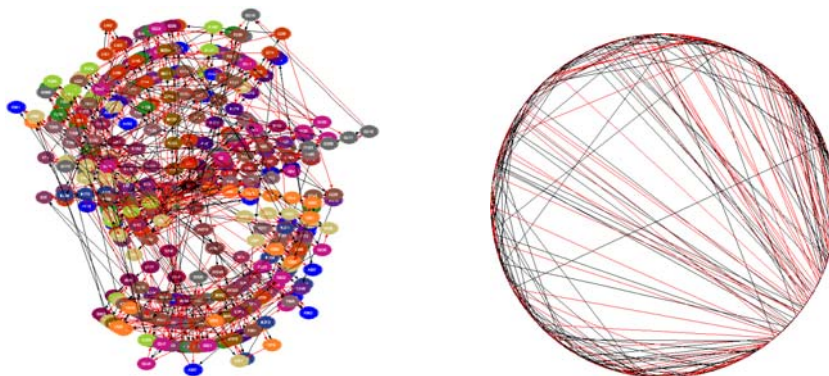


Figura 17: Ejemplo de resultados “embedded” con la S2SNet: índices topológicos y dibujos de los grafos de tipo estrella

El procesamiento de las secuencias se puede ver en una ventana con la consola. Si se cierra, todas las ventanas de la aplicación se cerrarán también. Los botones se pueden encontrar

también en el menú (sin el *Display*). Además, desde el menú se puede abrir el editor de texto Bloc de notas si necesita ver/editar sus archivos de entrada/salida o crear otros nuevos.

En el dibujo de los grafos, cada grupo tiene un color diferente. Si se quiere obtener dibujos diferentes, se pueden encontrar los archivos DOT (para cada secuencia) y los ejecutables del **Graphviz** (**dot**, **circo**, **twopi**, **neato**, **fdp**) en la carpeta “**dot**”.

El menú *Calculations* permite transformar sus datos en el formato S2SNet (una cadena de caracteres).

- **Numbers to Sequence** (Números en Secuencias) – Transforma los números (delimitados por TAB) en secuencias de caracteres; las opciones son las siguientes:
 - **Parameters** (Parámetros): los valores mínimos y máximos de los datos numéricos, el número de grupos que necesitamos (*number of groups*, hasta un máximo de 80); se puede utilizar el botón **GET** para emplear los valores mínimos y máximos calculados a partir de sus datos automáticamente
 - **Input files** (Archivos de entrada): archivo con datos como números
 - **Output files** (Archivos de salida): archivo con secuencias, archivo con grupos y archivo con intervalos de números (la descripción de los intervalos numéricos para cada grupo).

Nota: esta función se puede utilizar para transformar los valores de un espectro de masa proteica en secuencias para poder calcular los índices topológicos del grafo de tipo estrella.

- **N to 1-Character Sequence** – transforma las secuencias donde la información está codificada en N caracteres en secuencias tipo S2SNet basadas en cada carácter; las opciones son las siguientes:
 - **Input files:** archivos de entrada con secuencias codificadas en N-caracteres (*N-character file*) (archivo inicial), archivo con la codificación (*code file*), la equivalencia entre N-caracteres y 1-carácter (ex: ALA=A)
 - **Output files:** archivos de salida para secuencias típicas a S2SNet (*1-character*) (archivo final) y para los grupos (*group file*).

Nota: esta función se puede emplear para transformar las secuencias codificadas en 3 letras tales como los codones para los aminoácidos en secuencias de tipo S2SNet con aminoácidos como un carácter.

Índices de las redes de tipo estrella

Sus datos se utilizarán para calcular los siguientes índices:

✓ **Entropía de Shannon** de las n Matrices de Markov (Sh)

$$Sh_n = - \sum_i p_i * \log(p_i) \quad (11)$$

p_i son los elementos n_i del vector p resultado desde la multiplicación vectorial entre la matriz Markov normalizada ($n_i \times n_i$) elevada al poder y el vector ($n_i \times 1$) con cada elemento igual a $1/n_i$;

✓ **Traces** de las matrices de conectividad (Tr)

$$Tr_n = \sum_i (M^n)_{ii} \quad (12)$$

$n = 0 \dots$ poder, M = matriz conectividad (dimensión i^*i); $ii = i$ -ésimo elemento diagonal;

✓ Número de **Harary** (H)

$$H = \sum_{i < j} \frac{m_{ij}}{d_{ij}} * w_j^{nw} \quad (13)$$

d_{ij} = elementos de la matriz de distancia, m_{ij} = elementos de la matriz de conectividad M , w_j = los pesos, nw = es (1) para la selección de los pesos y (0) al contrario;

✓ Index de **Wiener** (W)

$$W = \sum_{i < j} d_{ij} * w_j^{nw} \quad (14)$$

✓ Índice Topológico de **Gutman** (S6)

$$S_6 = \sum_{ij} \frac{deg_i * deg_j}{d_{ij}} * w_j^{nw} \quad (15)$$

deg_i = elementos de la matriz de los grados;

✓ Índice Topológico de **Schultz** (S)

$$S = \sum_{i < j} (deg_i + deg_j) * d_{ij} * w_j^{nw} \quad (16)$$

✓ Índice de autocorrelación de la estructura topológica de **Moreau-Broto** (ATSn, $n=1 \dots$ poder); sólo si se incluyen los pesos:

$$ATSn = \sum_{ij} dp_{ij}^n * w_i * w_j \quad (17)$$

dp_{ij}^n = elementos de la matriz de distancias entre pares de nodos cuando la distancia es n ;

✓ Índice de conectividad de la distancia de **Balaban** (J)

$$J = \text{edges} / (\text{edges} - \text{nodes} + 2) * \sum_{i < j} m_{ij} \sqrt{\sum_k d_{ik} * \sum_k d_{kj}} * w_j^{nw} \quad (18)$$

nodes/edges = números de nodos/aristas en la red de tipo estrella;

✓ Índices de conectividad de **Kier-Hall**

$${}^0X = \sum_i \frac{w_i^{nw}}{\sqrt{\text{deg}_i}} \quad {}^2X = \sum_{i < j < k} \frac{m_{ij} * m_{jk} * w_k^{nw}}{\sqrt{\text{deg}_i * \text{deg}_j * \text{deg}_k}} \quad (19)$$

$${}^3X = \sum_{i < j < k < m} \frac{m_{ij} * m_{jk} * m_{km} * w_m^{nw}}{\sqrt{\text{deg}_i * \text{deg}_j * \text{deg}_k * \text{deg}_m}} \quad {}^4X = \sum_{i < j < k < m < o} \frac{m_{ij} * m_{jk} * m_{km} * m_{mo} * w_o^{nw}}{\sqrt{\text{deg}_i * \text{deg}_j * \text{deg}_k * \text{deg}_m * \text{deg}_o}} \quad (20)$$

$${}^5X = \sum_{i < j < k < m < o < q} \frac{m_{ij} * m_{jk} * m_{km} * m_{mo} * m_{oq} * w_q^{nw}}{\sqrt{\text{deg}_i * \text{deg}_j * \text{deg}_k * \text{deg}_m * \text{deg}_o * \text{deg}_q}} \quad (21)$$

✓ Índice de conectividad de **Randic**

$${}^1X = \sum_{i < j} \frac{m_{ij} * w_j^{nw}}{\sqrt{\text{deg}_i * \text{deg}_j}} \quad (22)$$

REGISTRO GENERAL DE LA PROPIEDAD INTELECTUAL

Según lo dispuesto en la Ley de Propiedad Intelectual (Real Decreto Legislativo 1/1996, de 12 de abril), quedan inscritos en este Registro los derechos de propiedad intelectual en la forma que se determina seguidamente:

NÚMERO DE ASIENTO REGISTRAL 03 / 2008 / 1338

Título: S2SNet - Sequence to Star Network

Objeto de propiedad intelectual: programa de ordenador

Clase de obra: programa de ordenador

PRIMERA INSCRIPCIÓN

Autorles y titularles originarios de derechos

- **Apellidos y nombre:** MUNTEANU, Cristian Robert
Nacionalidad: ROM **D.N.I./N.I.F./Pasaporte:** X-4541639-J
- **Apellidos y nombre:** GONZÁLEZ DÍAZ, Humberto
Nacionalidad: CUB **D.N.I./N.I.F./Pasaporte:** X-6672910-N

Datos de la solicitud

Núm. solicitud: SC-309-08

Fecha de presentación y efectos: 25/08/2008 **Hora:** 11:30

En Santiago de Compostela, a diecisiete de septiembre de dos mil ocho




José M^o Giljo Vázquez
Registrador



2.1.3. CULSPIN – Índices topológicos del grafo tipo espiral

2.1.3.1. Artículos publicados con grafos de tipo espiral

2.1.3.1.1. Clasificación cualitativa entre la estructura de las proteínas y el cáncer colorrectal utilizando las entropías tipo Shannon del grafo estrella y los métodos Naïve Bayes

Naïve Bayes QSDR classification based on spiral-graph Shannon entropies for protein biomarkers in human colon cancer

[Molecular BioSystems](#) 8, 1716-1722 (2012)

Vanessa Aguiar-Pulido, **Cristian Robert Munteanu**, José A Seoane, Enrique Fernández-Blanco, Lázaro G Pérez-Montoto, Humberto González-Díaz, Julian Dorado

Enlace: <http://goo.gl/JQQIE>

El diagnóstico rápido del cáncer representa una necesidad real en la medicina aplicada debido a la importancia de esta enfermedad. Los modelos teóricos pueden ayudar como herramientas de predicción. La representación teoría de grafos es una opción, ya que nos permite describir numéricamente cualquier sistema real, como las macromoléculas proteicas mediante la transformación de propiedades reales en índices topológicos de grafos moleculares. Este estudio propone un nuevo modelo de clasificación para las proteínas relacionadas con el cáncer de colon humano, mediante el uso de los índices topológicos del grafo tipo espiral sobre las secuencias de aminoácidos de proteínas. El mejor modelo cuantitativo de la relación estructura-enfermedad se basa en once índices de entropía de Shannon. Se obtiene con el método del clasificador bayesiano ingenuo (Naive Bayes) y muestra una excelente capacidad predictiva (90,92%) para nuevas proteínas vinculadas con este tipo de cáncer. El análisis estadístico confirma que este modelo permite el diagnóstico del cáncer de colon humano con AUROC de 0,91. La metodología que se presenta puede ser utilizada para cualquier tipo de información secuencial, como cualquier proteína o secuencias de ácidos nucleicos.

2.1.3.2. Manual del programa CULSPIN

¿Qué es CULSPIN?

CULSPIN (Compute ULam SPiral INdices) transforma cualquier secuencia de letras en una representación gráfica que usa como plantilla la espiral de Ulam (disposición de los números naturales en forma de espiral) y en la que se conectan aquellos nodos que pertenecen a la misma clase (tienen la misma letra). La interfaz se presenta en la **Figura 18**.

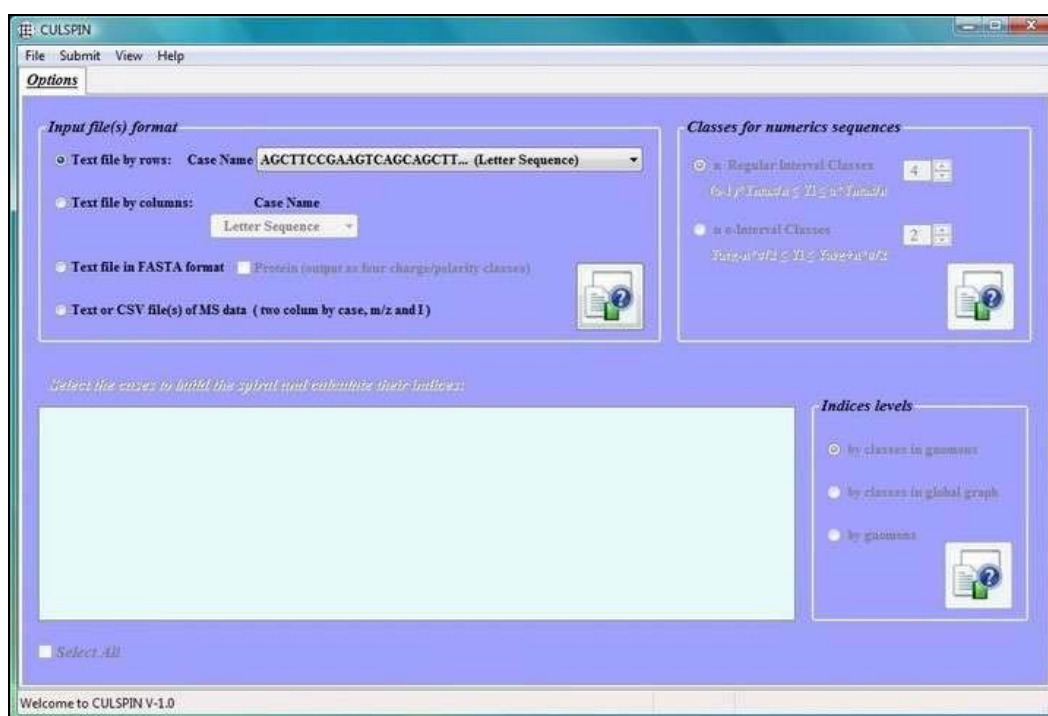


Figura 18: Interfaz del programa CULSPIN

Un ejemplo es el grafo tipo espiral en la **Figura 19** para la siguiente secuencia:

```
Cha[01]GDDGGDGGGGGGGDDGGGDGDDGGDGGGDGDDGGDGGDDDDGGGGGGDGGDDGGGGGG  
GGGGGGGGGGKKKKKAAAKKAKKKKKKAAA  
KKKKAKKKKAAAKKKKKKKKAAKKAATAAK
```

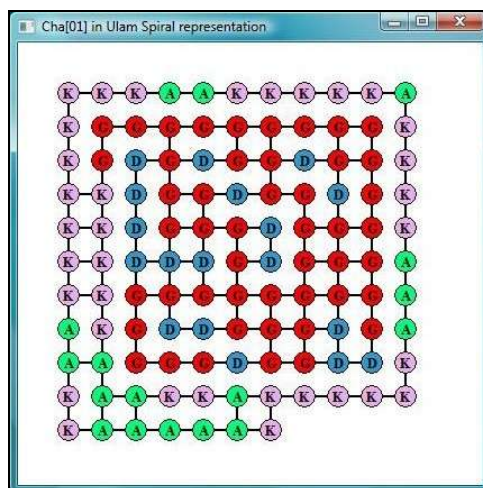


Figure 19: Grafo espiral para la secuencia Cha[01]

Además, basándose en este grafo, CULSPIN calcula dos familias de Índices Topológicos (*TIs*). Estos índices pueden ser calculados a varios niveles: para cada una las clases en cada gnomon de Ulam, para cada una de las clases en todo el grafo y para cada gnomon independiente de las clases. Por otra parte, los grafos 2D (*Grafos-U*) generados por la aplicación, además de ser visualizados, pueden ser exportados con el objetivo de poder utilizarlos en otros programas para calcular otras familias de *TIs*. Todos los índices numéricos se pueden guardar y/o exportar y con ellos se pueden realizar diversos análisis estadísticos o crear modelos QSAR (relación estructura-propiedades). Ejemplos de secuencias son las cadenas de aminoácidos de las proteínas, los ácidos nucleicos y los espectros de masas de las proteínas. CULSPIN se puede utilizar para estudiar distintos sistemas, desde los sistemas simples de átomos en pequeñas moléculas anti-cancerígenas, hasta sistemas complejos de redes metabólicas, sociales, computacionales o sistemas biológicos.

¿Qué puede hacer CULSPIN?

- ✓ *Leer* secuencias de letras organizadas en filas o columnas a partir de ficheros *TXT*;
- ✓ *Leer* secuencias en formato *FASTA* almacenadas en ficheros *TXT*;
- ✓ *Leer* secuencias o series numéricas, organizadas en filas o columnas a partir de ficheros *TXT*;
- ✓ *Leer* datos numéricos correspondientes a señales de Espectros de Masas (*MS*) a partir de múltiples ficheros *TXT* o *CSV*;

- ✓ **Convertir** secuencias o series numéricas y datos de *MS* en secuencias de letras;
- ✓ **Transformar** cualquier secuencia de letras en su correspondiente **Grafo-U** conectando los nodos que pertenezcan a la misma clase (tienen la misma letra);
- ✓ **Calcular** dos familias de *TIs* usando los **Grafos-U** generados y **Mostrar** sus valores en una tabla;
- ✓ **Graficar y Visualizar** el **Grafo-U** de la secuencia que se seleccione;
- ✓ **Exportar** la información de la conectividad de los **Grafo-U** en ficheros *CT* o *NET*;
- ✓ **Guardar** los *TIs* calculados en ficheros *TXT* o *CSV*.

¿Cómo utilizar CULSPIN?

CULSPIN es una aplicación interactiva creada con Python/wxPython con formato de libreta de notas que presenta una barra de menú principal con las siguientes opciones:

Menú *File*:

- **Open file:** permite buscar, seleccionar, abrir el fichero del cual se tomarán los datos de entrada (secuencias de letras, secuencias o series numéricas, etc.). Una vez cargados los datos, las secuencias de letras se muestran en una lista.
- **Reload sequences:** permite volver a trabajar con las secuencias cargadas inicialmente (secuencias originales). Esta opción sólo se activa si no se le construyó la espiral a todas las secuencias originales. Una vez terminado el proceso de recarga, todas las secuencias originales vuelven a estar disponibles en la lista.
- **Make a copy of:** hacer una copia en un fichero *TXT* de las secuencias de letras originales o las secuencias de letras estudiadas, pero en el formato con el que se muestran en la lista (**nombre <espacio>secuencia**). Esta opción está disponible o activa sólo cuando las secuencias mostradas al abrir el fichero, han requerido cierta transformación, es decir, cuando los datos estaban organizados por columnas, eran números, estaban en formato *FASTA*, etc.
- **Export graph:** exportar a ficheros independientes de tipo *CT* o *NET* la conectividad de cada uno de los **Grafos-U** construidos, con el objetivo de poder utilizarlos en otros programas para someterlos a otros cálculos.
- **Save Indices:** guardar en ficheros *TXT* o *CSV* los índices calculados por la aplicación para su posterior estudio estadístico

- **Quit:** salir de la aplicación.

Menú *Submit*:

- **Build Spiral:** colocar las secuencias seleccionadas en la representación de espiral y construir el **Grafo-U** conectando los nodos que pertenecen a la misma clase (los que tienen la misma letra).
- **Calculate Indices:** calcular los **TIs** de las secuencias seleccionadas a partir de sus respectivos **Grafos-U**. Una vez terminada esta operación, los resultados se muestran en una nueva página.

Menú *View*:

- **View a graph:** graficar y visualizar, en una ventana independiente, el **Grafo-U** de una secuencia seleccionada (una secuencia a la vez). Sólo está activa después de haber construido al menos un **Grafo-U**.

Menú *Help*:

- **Help:** muestra en una ventana independiente el contenido la ayuda.
- **About:** muestra la clásica ventana con información acerca de la aplicación.

En un inicio, CULSPIN presenta una sola página con el título **Options** en su ventana principal en forma de libreta de notas.

Página *Options*:

En esta página hay cuatro áreas bien definidas cuyas funciones se describen a continuación:

I-Input file(s) format: esta caja de controles permite seleccionar, entre los tipos de formatos de ficheros de entrada aceptados por CULSPIN, aquella opción que se corresponda con el formato de nuestros datos.

A continuación mostramos un ejemplo de cada uno de los formatos para su mejor comprensión.

Secuencias numéricas:

Cha[01] Cha[02] Cha[03]
-7.86E-05 2.18E-07 9.60E-05
2.18E-07 9.60E-05 0.00036601
9.60E-05 0.00036601 0.0008102
0.00036601 0.0008102 0.00142856
0.0008102 0.00142856 0.00222112
0.00142856 0.00222112 0.00318787
0.00222112 0.00318787 0.00432881
0.00318787 0.00432881 0.00564393
0.00432881 0.00564393 0.00713324
0.00564393 0.00713324 0.00879674
0.00713324 0.00879674 0.01063443
0.00879674 0.01063443 0.01264631
0.01063443 0.01264631 0.01483238
0.01264631 0.01483238 0.01719263
0.01483238 0.01719263 0.01972708

c-) Text file in FASTA format:

```
>gi|221068402|ref|ZP_03544507.1|enzyme [Comamonas testosteroni KF-1]
MSEPVNQWPQTLLEERIDRLES�DAIRQLAGKYSLSLDMRMDAHVNLFAPDIKVGKEKVGRAHFMAWQDS
TLRDQFTGTSHHLGQHIIIEFVDRDHATGVVYSKNEHECGAEWVIMQMLYWDDYERIDGQWYFRRRLPCYW
YATDLNKPPIGDMKMRWPGREPYHGAFHELFPWKEFWAQRPGKDQLPQVAAPAPLEQFLRTMRRGTPAP
RMRVR
```

```
>gi|220713425|gb|EED68793.1| enzyme [Comamonas testosteroni KF-1]
MSEPVNQWPQTLLEERIDRLES�DAIRQLAGKYSLSLDMRMDAHVNLFAPDIKVGKEKVGRAHFMAWQDS
TLRDQFTGTSHHLGQHIIIEFVDRDHATGVVYSKNEHECGAEWVIMQMLYWDDYERIDGQWYFRRRLPCYW
YATDLNKPPIGDMKMRWPGREPYHGAFHELFPWKEFWAQRPGKDQLPQVAAPAPLEQFLRTMRRGTPAP
RMRVR
```

```
>gi|77360245|ref|YP_339820.1| enzyme [Pseudoalteromonas haloplanktis TAC125]
MQYLVISDIYGKTPCLQLAKHFNAENQIVDPYNGVHQALENEEEYKLFIKHCGHDEYAAKLEEFNKL
SKPTICIAFSAGASAAWRAQASTTTTHLKKVIAFYPTQIRNYLNIDAIHPCEFIFPGFEPHFVDELITN
LSAKNNVRCLKTLYLHGFMNQSQNFSEYGYQYFYKVIKTANSEAH
```

Note: en el caso de las proteínas, si se selecciona la opción **Protein**, cada aminoácido presente en la secuencia se codifica con una letra o clase diferente. Para ello se tiene en cuenta el grupo al que pertenezca el aminoácido según la polaridad y las propiedades ácido-base de sus cadenas laterales: **no polar y neutro; polar y neutro; ácido y polar; y básico y**

polar.

d-) Text or CSV files of MS data: En esta opción cada caso se encuentra almacenado en un fichero independiente. En ellos los datos de las señales del espectro están organizados en dos columnas: *masa/carga (m/z)* e *Intensidad* con encabezado o no. Los ficheros pueden ser de tipo *TXT* o *CSV*.

Ficheros TXT: (las columnas están separadas por tabulación)

2.5660	0.6601
3.6601	8.9102
8.1024	42.0856
14.2856	22.2112
22.2112	3.8787
31.8787	4.3288
43.2881	56.4393
56.4393	71.3324
71.3324	87.9674
87.9674	90.0000
106.3443	12.1631
126.4631	8.3238
148.3238	100.9263

Ficheros CSV: (los elementos están separados por comas)

m/Z,Intensity
2.5660,0.6601
3.6601,8.9102
8.1024,42.0856
14.2856,22.2112
22.2112,3.8787
31.8787,4.3288
43.2881,56.4393
56.4393,71.3324
71.3324,87.9674
87.9674,90.0000
106.3443,12.1631
126.4631,8.3238
148.3238,100.9263

II-Classes for numerical sequences: esta caja de controles sólo está activa si el formato de entrada seleccionado es de tipo numérico. En ella se ofrecen dos heurísticas diferentes para transformar una secuencia o serie numérica en una secuencia de letras.

- **n Regular Interval Classes:** en esta opción los datos numéricos tomados del fichero de entrada se dividen en **n** intervalos o clases ($2 \leq n \leq 10$) y se les asigna una letra diferente. Entonces, cada elemento de la secuencia o serie numérica se codifica con la letra de la clase a la que pertenece.

- ***n* σ -Interval Classes:** en esta opción los datos numéricos tomados del fichero de entrada se dividen en $2n+2$ intervalos ($2 \leq n \leq 4$) cuyas dimensiones dependen de la desviación estándar de los datos. A cada intervalo o clase se le asigna una letra y se codifica cada elemento de la secuencia o serie numérica con la letra de la clase a la que pertenezca.

Note: En el caso de los datos de MS, la presente versión de CULSPIN, los transforma previamente en una serie numérica en la que cada elemento es el producto de la *m/z* por la intensidad de cada señal del espectro. Luego esta serie numérica es transformada en una secuencia de letras utilizando la heurística seleccionada por el usuario.

III-A list box for view/select sequences: esta caja de lista tiene la función de mostrar y permitir la selección de secuencias o casos (**Figura 21**). En un inicio la lista está vacía y después de leer los datos a partir del fichero de entrada, la lista muestra las secuencias leídas directamente del fichero u obtenidas mediante alguna codificación o transformación de las explicadas anteriormente. Una vez que las secuencias de letras son mostradas en esta caja de lista, aparece una invitación a seleccionar las secuencias o casos a los que se les desea construir su **Grafo-U**.

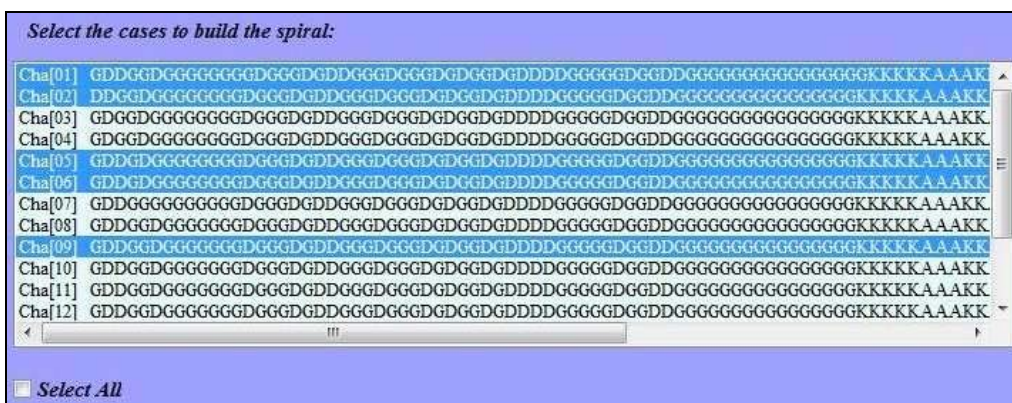


Figura 21. La selección de los casos para el CULSPIN

Se puede seleccionar un bloque continuo de secuencias o casos manteniendo presionada la tecla **Mayúsculas** al seleccionar el primero y el último caso que conforman el bloque; seleccionar casos alternos regularmente o no presionando la tecla **Ctrl** mientras se seleccionan los casos deseados; o seleccionar todos los casos marcando la opción **Select All**. Después de construir los **Grafos-U** de las secuencias seleccionadas, la caja de lista mostrará sólo los casos con los que se trabajó. En este momento se invita entonces a seleccionar los casos a los que se les desea calcular los **TIs** o seleccionar un único caso para ver su grafo en una ventana

independiente.

El resto de las secuencias no estudiadas se pueden recuperar sin necesidad de leer nuevamente el fichero de entrada, mediante la opción **Reload sequences** presente en el menú **File**. En tal caso se comienza desde cero, es decir, se perderán los grafos y los **TIs** calculados si no se han guardado en ficheros.

IV-Indices levels: esta caja de controles sólo se encuentra activa si se ha construido al menos una espiral y permite seleccionar a qué nivel queremos calcular las dos familias de **TIs** implementadas en esta versión de CULSPIN.

- **by classes in gnomons:** si se selecciona esta opción, las dos familias de **TIs** se calculan para cada una de las clases en cada uno de los gnómones. En el caso en que una clase no se encuentre en un determinado gnomon, su **Frecuencia** y su **Entropía de Shannon** en ese gnomon son cero. Esta opción es más útil cuando las secuencias no tienen muchas clases y no son muy grandes, en caso contrario, se obtendría un número demasiado elevado de índices y por tanto su procesamiento estadístico posterior muy engorroso.
- **by classes in global graph:** en esta opción los **TIs** se calculan para cada una de las clases pero en todo el grafo. En otras palabras, los **TIs** de una clase dada en todo el grafo, son el resultado de la sumatoria de sus valores en todos los gnómones. Esta opción reduce el número de **TIs** en el caso de secuencias muy grandes, por lo que resulta una buena opción en tales casos.
- **by gnomons:** si se selecciona esta opción, los **TIs** se calculan a nivel de gnómones independientemente de las clases. En otras palabras, los índices para un gnomon determinado son el resultado de la sumatoria de los **TIs** de todas las clases en ese gnomon. Esta opción puede ser muy útil si se trabaja con secuencias de tamaño moderado y con un gran número de clases.

Página Indices:

Esta página se adiciona a la libreta y se muestra al usuario inmediatamente después de que se calculen los **TIs** a las secuencias seleccionadas (**Figura 22**). El formato de la página es el de una tabla tipo hoja de cálculo, en la que en el encabezado de las columnas se muestran los nombres de los índices y el de las filas el de las secuencias o casos.

En esta tabla se puede seleccionar una celda, un rango, una columna, una fila o todas las celdas y copiar el contenido de la selección en el clipboard mediante la combinación **Ctrl+C** para luego pegarlo en donde se desee. Esta posibilidad es muy útil si se desea exportar de modo

rápido, sencillo y fácil los valores de los **TIs** calculados en aplicaciones externas tales como Excel.

	Fr(1)	Fr(2)	Fr(3)	Fr(4)	Fr(5)	Fr(6)	Fr(7)	Fr(8)	Fr(9)	Fr(10)	Sh(1)
Cha[01]	0.01645	0.03947	0.06579	0.07895	0.10526	0.10855	0.11513	0.13487	0.13158	0.14803	0.02934
Cha[02]	0.01020	0.04082	0.06122	0.08163	0.11224	0.10544	0.11905	0.13605	0.13265	0.15306	0.02032
Cha[03]	0.02000	0.04667	0.06333	0.08000	0.11000	0.10333	0.11667	0.13333	0.13000	0.15000	0.03398
Cha[04]	0.02000	0.04667	0.06333	0.08000	0.11000	0.10333	0.11667	0.13333	0.13000	0.15000	0.03398
Cha[05]	0.02000	0.04667	0.06000	0.08333	0.11000	0.10333	0.11667	0.13333	0.13000	0.15000	0.03398
Cha[06]	0.02000	0.04667	0.06000	0.08333	0.11000	0.10333	0.11667	0.13333	0.13000	0.15000	0.03398
Cha[07]	0.02000	0.04667	0.06000	0.08333	0.11000	0.10333	0.11667	0.13333	0.13000	0.15000	0.03398
Cha[08]	0.01689	0.04054	0.06081	0.08108	0.11149	0.10473	0.11824	0.13514	0.13176	0.15203	0.02994
Cha[09]	0.01689	0.04054	0.06081	0.08108	0.11149	0.10473	0.11824	0.13514	0.13176	0.15203	0.02994
Cha[10]	0.01689	0.04054	0.06081	0.08108	0.11149	0.10473	0.11824	0.13514	0.13176	0.15203	0.02994
Cha[11]	0.01689	0.04054	0.06081	0.08108	0.11149	0.10473	0.11824	0.13514	0.13176	0.15203	0.02994
Cha[12]	0.01689	0.04054	0.06081	0.08108	0.11149	0.10473	0.11824	0.13514	0.13176	0.15203	0.02994
Cha[13]	0.01689	0.04054	0.06081	0.08108	0.11149	0.10473	0.11824	0.13514	0.13176	0.15203	0.02994
Cha[14]	0.01689	0.04054	0.06081	0.08108	0.11149	0.10473	0.11824	0.13514	0.13176	0.15203	0.02994
Cha[15]	0.01689	0.04054	0.06081	0.08108	0.11149	0.10473	0.11824	0.13514	0.13176	0.15203	0.02994

Figura 22: Resultado con los TIs calculados en el CULSPIN

Espiral de Ulam

En 1963 el matemático *Stanisław M. Ulam* descubrió ciertos aspectos interesantes relacionados con la disposición que adoptan los números primos al colocar los números naturales en forma de una espiral. Luego esta disposición tomó mucho auge en la generación y visualización de imágenes.

Para construir la espiral se colocan los números en una rejilla de cuadrículas comenzando por 1 en el centro y luego los demás formando una espiral cuadrada según la **Figura 23**.

En matemáticas, esta representación es un método simple de graficar números con el que se revelen aspectos ocultos y muy interesantes de las series y secuencias numéricas. En el estudio de las moléculas, esta representación en espiral ha sido asociada en muchos trabajos encaminados a representar secuencias de nucleótidos de ADN divididos en cuatro clases (A,T,G y C).

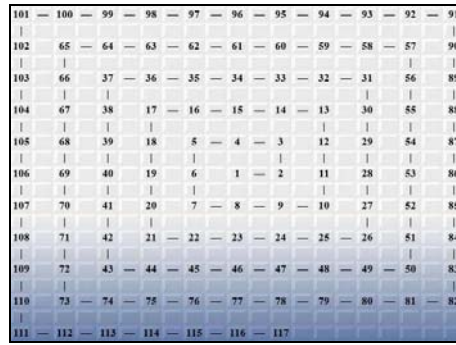


Figura 23: Espiral cuadrada con los datos

¿Qué es un gnomon?

La espiral de Ulam puede dividirse en diferentes regiones o intervalos nombrados gnómones o disposiciones angulares según se puede observar en la **Figura 24**. Para definir un gnomon es necesario recordar los números oblongos que son aquellos que se pueden representar mediante el producto $n(n+1)$ con n natural, es decir: 2, 6, 12, 20, 30, 42, 56, 72, 90,.... Estos números dividen a los números naturales en intervalos crecientes en longitud ($2n$). Resulta fácil de ver que un par de números oblongos consecutivos definen un gnomon y que estas disposiciones angulares se van encajando dando lugar a rectángulos de magnitud creciente. Además queda claro que cada elemento de la espiral pertenece a un único gnomon, es por ello que se puede definir la coordenada U de un elemento en la espiral de Ulam como el número del gnomon al que pertenece.

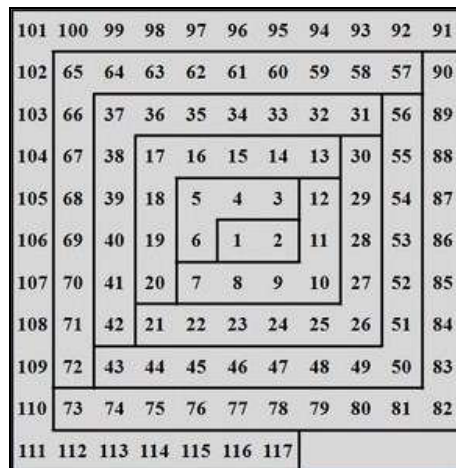


Figura 24: Representación de números por gnómones de un grafo espiral de Ulam

Cuando se representa una secuencia de letras en su *Grafo-U*, cada nodo es un elemento

de la secuencia cuya letra representa la clase a la que pertenece dicho elemento y en cada gnomon existirán una o más clases diferentes (**Figura 25**).

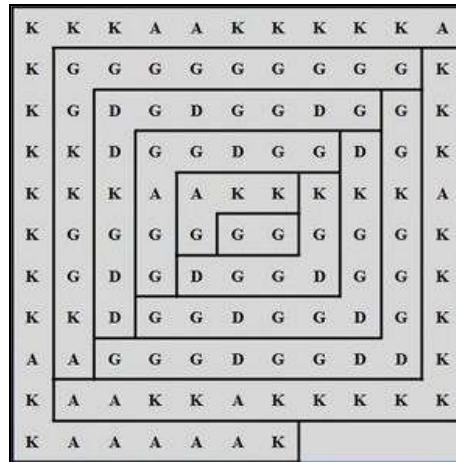


Figura 25: Representación de una secuencia de letras en gnómones

Índices, definición y cálculo

Como se ha comentado desde un inicio, en los **Grafos-U** construidos con ayuda de CULSPIN, cada nodo pertenece a una clase determinada y ellos no sólo están conectados siguiendo la secuencia de letras, sino que además aquellos nodos que pertenecen a la misma clase (tienen igual letra) se conectan entre sí. De modo que, en nuestros **Grafos-U** cada nodo estará conectado con uno o más nodos. Por definición, se conoce como **grados** de un nodo al número de nodos con los que está conectado el nodo en cuestión y por **grados totales** de un grafo a la suma de los grados de todos los nodos que conforman el grafo, entonces podemos definir como grados de un gnomon a la suma de los grados de los nodos que pertenecen a dicho gnomon. Teniendo en cuenta todo lo anterior, los índices calculados por CULSPIN se definen y calculan en las formulas desde la **Figura 26**.

Indices levels	Frequency	Shannon Entropy
by classes in gnomon	$f(c, g) = \frac{\sum_{n \in G_g} \deg(n(c, g))}{\sum_{i \in G_g} \deg(i)}$ <i>c: class; g: gnomon; $n_{c,g}$: node with class c in gnomon g</i>	$Sh(c, g) = -f(c, g) \log(f(c, g))$
by classes in global graph	$f(c) = \frac{\sum_{n \in G} \deg(n(c))}{\sum_{i \in G} \deg(i)}$ <i>c: class; n_c: node with class c in G_V</i>	$Sh(c) = -f(c) \log(f(c))$
by gnomons	$f(g) = \frac{\sum_{n \in G_g} \deg(n(g))}{\sum_{i \in G_g} \deg(i)}$ <i>g: gnomon; n_g: node in gnomon g</i>	$Sh(g) = -f(g) \log(f(g))$

Figura 26: Las formulas para los cálculos de gnómones

REGISTRO GENERAL DE LA PROPIEDAD INTELECTUAL

Según lo dispuesto en la Ley de Propiedad Intelectual (Real Decreto Legislativo 1/1996, de 12 de abril), quedan inscritos en este Registro los derechos de propiedad intelectual en la forma que se determina seguidamente:

NÚMERO DE ASIENTO REGISTRAL 03 / 2009 / 1199

Título: CULSPIN-Compute ULam SPiral INdices

Objeto de propiedad intelectual: programa de ordenador

Clase de obra: programa de ordenador

PRIMERA INSCRIPCIÓN

Autorles y titulares originarios de derechos

- **Apellidos y nombre:** PÉREZ MONTOTO, Lázaro Guillermo
Nacionalidad: CUB **D.N.I./N.I.F./Pasaporte:** X5119731-T
- **Apellidos y nombre:** PRADO PRADO, Francisco Javier
Nacionalidad: ESP **D.N.I./N.I.F./Pasaporte:** 44449687-W
- **Apellidos y nombre:** GONZÁLEZ DÍAZ, Humberto
Nacionalidad: CUB **D.N.I./N.I.F./Pasaporte:** X-6672910-N
- **Apellidos y nombre:** MUNTEANU, Cristian Robert
Nacionalidad: ROM **D.N.I./N.I.F./Pasaporte:** X-4541639-J

Datos de la solicitud

Núm. solicitud: SC-207-09

Fecha de presentación y efectos: 24/06/2009 **Hora:** 11:40

En Santiago de Compostela, a cuatro de septiembre de dos mil nueve




José M. Guijo Vázquez
REGISTRADOR DA PROPIEDAD INTELECTUAL



2.2. Nuevos servidores online Bio-AIMS basados en técnicas de ingeniería informática e inteligencia artificial

The screenshot displays the TargetPred @ Bio-AIMS website. At the top left is the logo for the Ibero-NBIC Network, RNASA-IMEDI4, TIC, Computer Science Faculty, University of A Coruña, Spain. The main title is "TargetPred @ Bio-AIMS" with the tagline "# Modelling the reality". On the top right is a logo with a pink and purple circular design and the text "Home | Links | About". Below the header, a description states: "Target Prediction: applications for predicting the function of several targets such as proteins in human diseases or molecular processes by using data such as protein sequences or blood proteome mass spectra". The main content area features a grid of tool icons and descriptions:

- EnzClassPred**: Enzyme Class Prediction
- ATCUNpred**: ATCUN DNA-cleavage protein activity Prediction
- Trypano-PPI**: Trypanosome Protein-Protein Interactions
- Plasmod-PPI**: Plasmodium Protein-Protein Interactions
- NL-MIND-BEST**: Non-Linear MARCH-INSIDE Nested Drug-Bank Exploration & Screening Tool
- MISSProt-HP**: MARCH-INSIDE Spectral moment prediction of Self Proteins in Human Parasites (other than original source organism)
- MIND-BEST**: Linear MARCH-INSIDE Nested Drug-Bank Exploration & Screening Tool
- LIBPpred**: Lipid-Binding Proteins Prediction
- LectinPred**: Lectin Prediction

Figura 27: El portal online **Bio-AIMS** (TargetPred sección) con las nuevas herramientas informáticas

Bio-AIMS (<http://bio-aims.udc.es/>) es una colección de servidores online que ofrece modelos teóricos basados en la Inteligencia Artificial, Biología Computacional y Bioinformática para estudiar sistemas complejos en ciencias ómicas (genómica transcriptómica, metabolómica, reactómica), que son relevantes en Parasitología, Microbiología, el cáncer, neurociencias, enfermedades cardiovasculares y otras investigaciones biomédicas en general. Los modelos se basan en los programas de ordenador **MARCH-INSIDE**, **MInD-Prot**, **S2SNet** y **MCeCoNet**. Es el resultado de la colaboración de dos grupos de la Red Gallega de Bioinformática (**RGB**):

Departamento de Tecnologías de la Información y las Comunicaciones (TIC), Facultad de Informática, Universidad de A Coruña ([UDC](#)), España y el Departamento de Microbiología y Parasitología de la [Facultad de Farmacia](#), Universidad de Santiago de Compostela (USC), España.

Bio-AIMS está dividido en dos tipos de herramientas:

- 1) **TargetPred - Target Prediction (Figure 27)** – Predicción de dianas: aplicaciones web para predecir la función de dianas diversas tales como las proteínas en enfermedades humanas o procesos moleculares, utilizando información a partir de las secuencias proteicas o la estructura 3D de las proteínas y a partir de la estructura química de los fármacos (SMILES)
- 2) **DiseasePred - Disease Prediction** – Predicción de enfermedades: aplicaciones en Biomedicina que ayudan en la predicción de enfermedades humanas utilizando datos biológicos tales como las mutaciones genéticas tipo *Single Nucleotide Polymorphism* (SNP), registros de EEG o espectros de masas del proteoma de la sangre

Los servidores presentados en esta tesis están dentro de la sección sobre el *TargetPred*: *Trypano-PPI*, *Plasmod-PPI*, *ATCUNpred* y *LectinPred*. Desde el 11 de febrero de 2010 hasta el 20 de abril de 2013, el servidor ha tenido más de 5000 visitas únicas desde 101 países (**Figura 28**).



Figura 28: Mas de 5000 visitas de las herramientas online del Bio-AIMS desde el 11 de febrero 2010 hasta el 20 de abril de 2013

2.2.1. Trypano-PPI – Interacciones proteína-proteína en *Trypanosoma*

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(2), 1182–1190 (2010)

Yamilet Rodríguez-Soca, **Cristian R. Munteanu**, Julián Dorado, Alejandro Pazos, Francisco J. Prado-Prado, and Humberto González-Díaz

Enlace: <http://goo.gl/nCgR9>

Herramienta: <http://bio-aims.udc.es/TrypanoPPI.php>

Ibero-NBIC Network
RNASA-IMEDIR, TIC
Computer Science
Faculty
University of A Coruña
Spain

TrypanoPPI @ Bio-AIMS

Modelling the reality

Home | Links | About

Trypanosome

Trypano-PPI
Trypanosome
Protein-Protein
Interactions (TPPI)

Tool: MARCH-INSIDE
(Python version)
Data: RCSB PDB

PDB-chain lists : Please paste the names of the PDB chains as two lists (maximum 50)
Notes: There is no space between the PDB name and the chain label, no empty new line; the results will print the combination between the chains from the first list and the chains from the second one.

1HOZA
1K3TB

1HOZB
1F2CA

Predict

LNN 2:2-1:1
(programed by Humberto González-Díaz)

Test Accuracy = 90.9%
(Training Accuracy = 89.5%)

Figura 29: Herramienta online TrypanoPPI

Trypanosoma brucei causa la tripanosomiasis africana en los seres humanos (HAT o enfermedad del sueño africano) y Nagana en el ganado. La enfermedad amenaza a más de 60

millones de personas y la innumerable cantidad de ganado en 36 países de África subsahariana, teniendo un impacto devastador en la salud humana y en la economía. Por otro lado, el *Trypanosoma cruzi* es el responsable en América del Sur por la enfermedad de Chagas, que puede causar una enfermedad grave y muerte, especialmente en niños pequeños. En este contexto, el descubrimiento de dianas terapéuticas nuevas en *Trypanosoma proteoma* es muy importante para la comunidad científica.

Recientemente, muchos investigadores han dedicado importantes esfuerzos en el estudio de las interacciones proteína-proteína (PPIs = Protein-Protein Interactions) en las especies patógenas de *Trypanosoma* y concluyeron que la identidad baja entre algunas proteínas de parásitos y su huésped humano convierten a estas PPIs en dianas farmacológicas muy prometedoras. No hay modelos generales conocidos para predecir PPIs únicas en *Trypanosoma* (TPPIs).

Por otro lado, la estructura 3D de un número creciente de proteínas de *Trypanosoma* se encuentra en las bases de datos. En este sentido es muy importante la introducción de un nuevo modelo para predecir el TPPI de la estructura 3D de proteínas implicadas en las PPI. Por eso, hemos introducido nuevos invariantes de los complejos proteína-proteína basados en el potencial electrostático Markov promedio $\xi_k(Ri)$ para de los aminoácidos ubicados en diferentes regiones (Ri) de la proteína i -ésima y colocada a una distancia k una de la otra. Se calcularon más de 30 tipos diferentes de parámetros para 7866 pares de proteínas (1023 TPPIs y 6823 no TPPI) de más de 20 organismos, incluyendo parásitos y huéspedes humanos o bovinos. Hemos encontrado un modelo lineal simple que predice más del 90% de los TPPIs y no TPPIs tanto en el entrenamiento y como en el grupo de validación utilizando sólo dos parámetros. Los parámetros son $d\xi_k(s) = |\xi_k(s1) - \xi_k(s2)|$, la diferencia absoluta entre los valores $\xi_k(s_i)$ en la superficie de las dos proteínas de los pares. También hemos probado los modelos no lineales tipo ANN con fines de comparación, pero el modelo lineal da mejores resultados. Hemos implementado este modelo en el servidor Web denominado TrypanoPPI, a la disposición del público de forma gratuita en <http://bio-aims.udc.es/TrypanoPPI.php> (**Figura 29**). Este es el primer modelo que predice si los complejos proteína-proteína en el proteoma de *Trypanosoma* son únicos con respecto a otros parásitos y huéspedes, abriendo nuevas oportunidades para el descubrimiento de dianas para fármacos anti-*Trypanosoma*. Un ejemplo de resultado para los pares entre las listas de cadenas proteicas [1HOZA, 1K3TB] y [1HOZB, 1F2CA] se presenta en la **Figura 30**.



Process ID = 109305174346d783d5
PDB List 1 = 1HOZA 1K3TB
PDB List 2 = 1HOZB 1F2CA

... please wait ...

PDB Update/Verification [List 1] ...
1HOZA 1K3TB

PDB Update/Verification [List 2] ...
1HOZB 1F2CA

Processing PDB-chain List 1 ...
1HOZA 1K3TB

Processing PDB-chain List 2 ...
1HOZB 1F2CA

Result file = Results/109305174346d783d5/TrypanoPPI.calc.txt

TrypanoPPI @ Bio-AIMS

Biopython server to predict if a pair of proteins form a physically stable complex unique of Trypanosoma (not present in human or other parasites) based on electrostatic potential indices of Protein-Protein Interactions (PPIs) by using MARCH-INSIDE (Python version) and LNN 2:2-1:1 (90.9% accuracy).
These complexes may be interesting candidates for specific anti-Trypanosoma drug targets.

Results = <http://bio-aims.udc.es/Results/109305174346d783d5/TrypanoPPI.calc.txt>
Calculated at 2013-04-21 20:48:14

Chain1	Chain2	Complex
1HOZA	1HOZB	YES
1HOZA	1F2CA	NO
1K3TB	1HOZB	NO
1K3TB	1F2CA	NO

Done!

Figura 30: Ejemplo de cálculo con el servidor TrypanoPPI

2.2.2. Plasmod-PPI – Interacciones proteína-proteína en *Plasmodium*

Plasmod-PPI: a web-server predicting complex biopolymer targets in Plasmodium with entropy measures of protein-protein interactions

[Polymer](#) 51(1), 264-273 (2010)

Yamilet Rodriguez-Soca, **Cristian R. Munteanu**, Julian Dorado, Juan Rabuñal, Alejandro Pazos and Humberto González-Díaz

Enlace: <http://goo.gl/hRhm9>

Herramienta: <http://bio-aims.udc.es/PlasmodPPI.php>

Ibero-NBIC Network
RNASA-IMEDIR, TIC
Computer Science Faculty
University of A Coruña
Spain

PlasmodPPI @ Bio-AIMS
Modelling the reality

Home | Links | About

Plasmod-PPI
Plasmodium Protein-Protein Interactions (PPPI)
Tool: MARCH-INSIDE (Python version)
Data: RCSB PDB

PDB-chain lists : Please paste the names of the PDB chains as two lists (max. 50)
Notes: There is no space between the PDB name and the chain label, no empty new line; the results will print the pairs between the chain from list 1 with the chain from list 2 (not the combination of the list items).

3C5IA
2F6IE
1SYRC

3C5IE
2GHUA
1SYRF

Classification Tree
Test Accuracy = 96.8 %

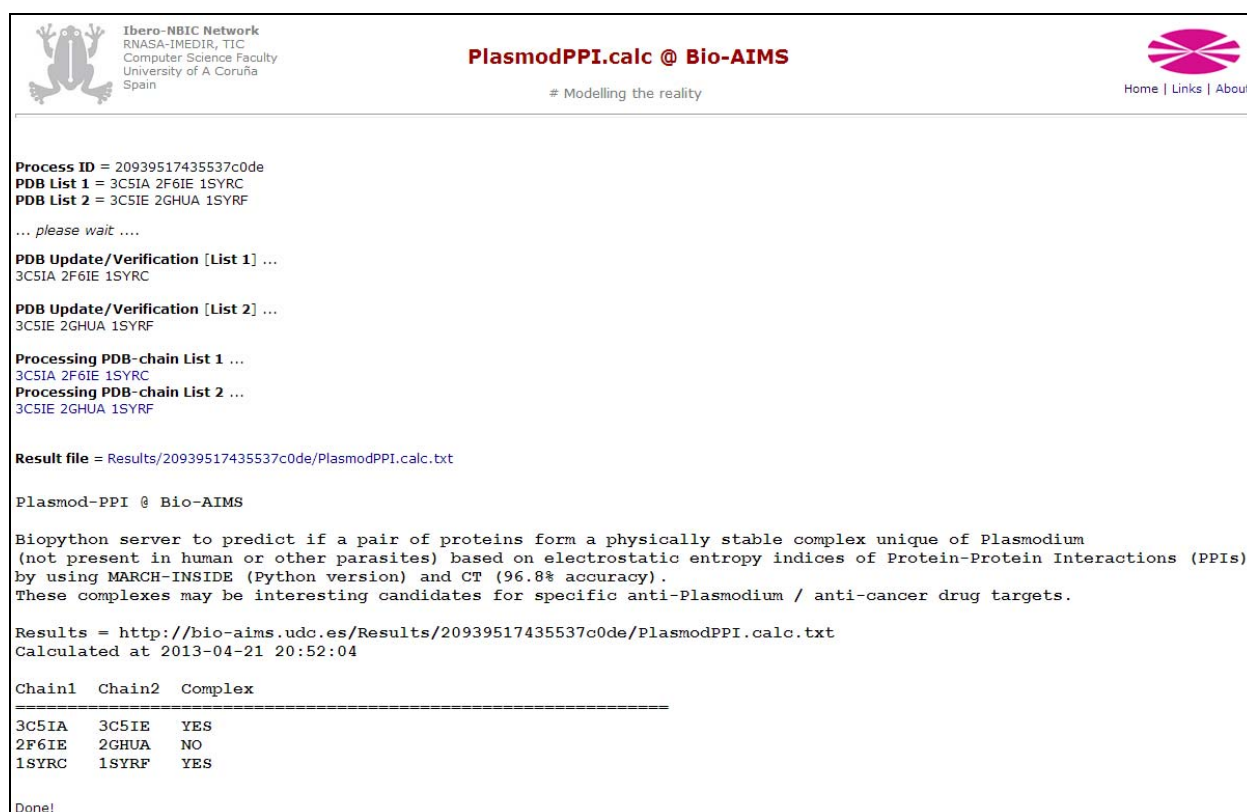
Predict

Figura 31: Herramienta online PlasmodPPI

Podemos definir los índices estructurales de polímeros o biopolímeros complejos, y usarlos en la predicción de nuevos fármacos y sus correspondientes dianas en los parásitos. Por ejemplo, el *Plasmodium falciparum* produce la forma más severa de malaria y mata hasta 2,7 millones de personas anualmente, mientras que *Plasmodium vivax* es geográficamente la causa con más distribución, con más de 80 millones de casos clínicos. Debido a la farmacorresistencia y la toxicidad, el descubrimiento de nuevas dianas de fármacos es obligatorio, tales como los complejos proteína-proteína únicos de este patógeno, pero no en el huésped humano (pPPCs).

Además, la estructura 3D de un número creciente de proteínas de *Plasmodium* se está introduciendo en las bases de datos públicas, facilitando el desarrollo de modelos bioinformáticos para predecir pPPCs. Además, algunos PPCs se expresan en los parásitos y en los humanos, tales como la DHFR sintetasa, juegan un papel importante en la resistencia a los medicamentos, tanto en la malaria como en el cáncer humano.

Sin embargo, no hay modelos generales para predecir los pPPCs utilizando los índices de la estructura del biopolímero PPC. Por lo tanto, en este trabajo presentamos nuevos descriptores numéricos de la cadena de Markov para las interacciones proteína-proteína (PPIs), basados en la entropía electrostática y se calculan estos parámetros para 5257 pares de proteínas (774 pPPCs y 4483 no pPPCs) de más de 20 organismos, incluyendo parásitos y huéspedes humanos. Se encontró un árbol de clasificación simple, con una alta precisión, sensibilidad y especificidad (90,2 - 98,5%), tanto en el entrenamiento como en la validación y se implementó en el servidor PlasmodPPI, fácil de usar, disponible de forma gratuita en <http://bio-aims.udc.es/PlasmodPPI.php> (Figura 31). Un ejemplo de resultado para los pares entre las listas de cadenas proteicas [3C5IA, 2F6IE, 1SYRC] y [3C5IE, 2GHUA, 1SYRF] se presenta en la Figura 32.



```

Ibero-NBIC Network
RNASA-IMEDIR, TIC
Computer Science Faculty
University of A Coruña
Spain

PlasmodPPI.calc @ Bio-AIMS
# Modelling the reality

Home | Links | About

Process ID = 20939517435537c0de
PDB List 1 = 3C5IA 2F6IE 1SYRC
PDB List 2 = 3C5IE 2GHUA 1SYRF

... please wait ...

PDB Update/Verification [List 1] ...
3C5IA 2F6IE 1SYRC

PDB Update/Verification [List 2] ...
3C5IE 2GHUA 1SYRF

Processing PDB-chain List 1 ...
3C5IA 2F6IE 1SYRC
Processing PDB-chain List 2 ...
3C5IE 2GHUA 1SYRF

Result file = Results/20939517435537c0de/PlasmodPPI.calc.txt

Plasmod-PPI @ Bio-AIMS

Biopython server to predict if a pair of proteins form a physically stable complex unique of Plasmodium
(not present in human or other parasites) based on electrostatic entropy indices of Protein-Protein Interactions (PPIs)
by using MARCH-INSIDE (Python version) and CT (96.8% accuracy) .
These complexes may be interesting candidates for specific anti-Plasmodium / anti-cancer drug targets.

Results = http://bio-aims.udc.es/Results/20939517435537c0de/PlasmodPPI.calc.txt
Calculated at 2013-04-21 20:52:04

Chain1 Chain2 Complex
=====
3C5IA 3C5IE YES
2F6IE 2GHUA NO
1SYRC 1SYRF YES

Done!

```

Figura 32: Ejemplo de cálculo con el servidor PlasmodPPI

2.2.3. ATCUNpred – Predicción de dianas proteicas con actividad ATCUN en parasitos

Complex Network Spectral Moments for ATCUN Motif DNA Cleavage: First Predictive Study on Proteins of Human Pathogen Parasites

[Journal of Proteome Research](#) 8(11), 5219–5228 (2009)

Cristian R Munteanu, José M. Vázquez, Julián Dorado, Alejandro Pazos Sierra, Ángeles Sánchez-González, Francisco J. Prado-Prado and Humberto González-Díaz

Enlace: <http://goo.gl/u7Thg>

Herramienta: <http://bio-aims.udc.es/ATCUNPred.php>

Ibero-NBIC Network
RNASA-IMEDIR, TIC
Computer Science Faculty
University of A Coruña
Spain

ATCUNPred @ Bio-AIMS
Modelling the reality

Home | Links | About

ATCUNpred
ATCUN DNA-cleavage protein activity
Prediction

Tool: MARCH-INSIDE (Python version)
Data: RCSB PDB

PDB list : Please paste the names of the PDB as a list (maximum 50)

1AZP
1I4M
1B0U

Predict

LDA classification model
Accuracy = 91.32%

$$DNA-cleavage = c_0 + \sum_{k=1}^n c_k * \pi_k(O)$$

$pi = electrostatic\ spectral\ moments$
 $O = amino\ acid\ orbits$

Note: For the sake of simplicity, in order to avoid the calculation of Mahalanobis distance, an approximate classification percentage is calculated as $(sc1)^2 / ((sc1)^2 + (sc2)^2)$, where $sc1$ and $sc2$ are the ATCUN and non-ATCUN scores (*DNA-cleavage*).

Figura 33: Herramienta online ATCUNPred

El desarrollo de métodos que pueden predecir la actividad biológica mediada del metal basado sólo en la estructura 3D de las proteínas no enlazadas con el metal se ha convertido en un objetivo de gran importancia. Este trabajo está dedicado a los motivos tipo terminal amino Cu(II) y Ni(II)-binding (ATCUN) que participan en la división del ADN y tienen actividad antitumoral.

Hemos calculado aquí, por primera vez, los momentos espectrales electrostáticos para la información proteica 3D de 415 proteínas diferentes, incluyendo 133 posibles proteínas ATCUN antitumoral. Utilizando estos parámetros como entrada para el análisis discriminante lineal, hemos encontrado un modelo que discrimina entre las proteínas de división ADN ATCUN y proteínas no activas con una precisión del 91,32% (379 de 415 de las proteínas que incluyen tanto el entrenamiento como la serie de validación externa).

Finalmente, el modelo ha predicho por primera vez la función de división del ADN de las proteínas de los parásitos patógenos. Nosotros hemos predicho posibles proteínas con actividad ATCUN, con una probabilidad superior al 99% en nueve familias de parásitos como *Trypanosoma*, *Plasmodium*, *Leishmania*, o *Toxoplasma*. La distribución de las funciones biológicas de las proteínas ATCUN predichas ha sido la siguiente: oxidorreductasas 70,5%, proteínas de señalización 62,5%, liasas 58,2%, proteínas de la membrana 45,5%, ligasas 44,4%, hidrolasas 41,3%, transferasas 39,2%, proteínas de adhesión celular 34,5%, *metal binders* 33,5%, proteínas de traducción 25,0%, proteínas de transporte molecular 16,7%, proteínas estructurales 9,1% y isomerasas 8,2%.

El modelo está implementado en <http://bio-aims.udc.es/ATCUNPred.php> (Figura 33). Un ejemplo de resultado para las proteínas 1AZP, 1I4M, 1B0U es:

```
ATCUNpred @ Bio-AIMS
ATCUN DNA-cleavage protein activity Prediction
by using MARCH-INSIDE and LDA based on electrostatic spectral moments
(Accuracy of 91.32%)
Results=http://bio-aims.udc.es/Results/24293517444a6099d0/ATCUNpred.calc2.txt
2013-04-21 21:57:26
```

```
PDB   ATCUN Prediction
=====
1AZP  0.28 %
1I4M  66.01 %
1B0U  80.97 %
```

2.2.4. LIBPpred – Predicción de proteínas que interacciona con los lípidos

LIBP-Pred: Web Server for Lipid Binding Proteins using Structural Network Parameters; PDB Mining of Human Cancer Biomarkers and Drug Targets in Parasites and Bacteria
[Molecular BioSystems](#) 8(3), 851-862 (2012)

Humberto González-Díaz, **Cristian R. Munteanu**, Lucian Postelnicu, Francisco Prado-Prado, Marcos Gestal, Alejandro Pazos

Enlace: <http://goo.gl/cTNcP>

Herramienta: <http://bio-aims.udc.es/LIBPpred.php>

Ibero-NBIC Network
RNASA-IMEDIR, TIC
Computer Science Faculty
University of A Coruña
Spain

LIBPpred @ Bio-AIMS
Modelling the reality

Home | Links | About

LIBPpred
Lipid-Binding Proteins Prediction
Tool: MARCH-INSIDE (Python version)

Mode 1: Standard PDBs
PDB/PDB chain List : Please paste the ID of the PDBs/PDB chains as a list
(maximum 10 items)

1QGK
1T4M
2QZTB
1B0U

Predict

Data: RCSB PDB

Mode 2: LOMETS PDB
Upload & evaluate one PDB from LOMETS (max. 2MB)
Please select LOMETS PDB

Seleccionar archivo No se ha seleccionado ningún archivo Predict

LDA classification model
Accuracy = **89.11%**
(the model is based on 9 spectral moments of the proteins)

Note: The LIBP prediction is calculated using
 $(LIBPscore - Min_score) * 100 / (Max_score - Min_score)$,
where LIBPscore is the result of the LDA equation for the current protein and Min and Max score are the minimum and maximum values of the LIBPscore for our dataset.

Figura 34: Herramienta online LIBPpred

Las proteínas que se unen a lípidos (Lipid-Binding Proteins, LIBPs) o proteínas de unión a los ácidos grasos (Fatty Acid-Binding Proteins, FABPs) juegan un papel importante en muchas enfermedades, tales como diferentes tipos de cáncer, lesión renal, aterosclerosis, diabetes, isquemia intestinal e infecciones parasitarias. Por lo tanto, los métodos computacionales que

pueden predecir LIBPs basado en parámetros de la estructura 3D se convirtieron en un objetivo de gran importancia para el descubrimiento de fármacos y sus dianas moleculares y para el diseño de vacunas y la selección de biomarcadores. El banco de datos de proteínas (PDB) contiene 3000 estructuras 3D de proteínas con función desconocida. Esta lista, así como los últimos resultados experimentales en la investigación proteómica, es una fuente muy interesante para descubrir proteínas relevantes, incluyendo LIBPs. Sin embargo, no hay modelos generales para predecir nuevos LIBPs basados en estructuras 3D. Se han desarrollado nuevos modelos de relaciones cuantitativas estructura-actividad (QSAR) en base a los parámetros electrostáticos 3D utilizando 1801 proteínas diferentes, incluyendo 801 LIBPs. Se calcularon los parámetros electrostáticos con la herramienta MARCH-INSIDE que se corresponden con la proteína entera o con regiones específicas de las proteínas: núcleo, interna, media y superficie (core, inner, middle, surface). Se utilizan estos parámetros como entradas para alimentar a un clasificador de análisis discriminante lineal (Linear Discriminant Analysis, LDA), que discriminará las estructuras 3D de los LIBPs de nuevas proteínas. Se implementa este predictor y se pone disponible gratuitamente en el servidor “Web” denominado LIBP-Pred, <http://bio-aims.udc.es/LIBPpred.php> (**Figura 34**). Los usuarios pueden realizar una recuperación automática de estructuras de proteínas desde PDB Web site o cargar sus modelos estructurales de proteínas personalizadas de su computador a través del servidor LOMETS. Se ha demostrado la posibilidad de efectuar un estudio predictivo de aproximadamente 2000 proteínas con función desconocida. Se han obtenido resultados interesantes con respecto al descubrimiento de nuevos biomarcadores de cáncer en los seres humanos o las dianas de fármacos antiparasitarios. Un ejemplo de resultado para las proteínas/cadenas proteicas 1QGHK, 1I4M, 2QZTB, 1B0U se presenta en la **Figura 35**.



Process ID = 17562517431e642d57

... please wait

PDB Update/Verification ...
1QG HK 1I4M 2QZTB 1B0U Done!

Calculating ...

Result file = Results/17562517431e642d57/LIBPpred.Mode1.txt

LIBPpred @ Bio-AIMS

Mode 1: Standard PDB input

Lipid-Binding Proteins Prediction

by using MARCH-INSIDE and LDA based on electrostatic spectral moments (Accuracy of 89.11%)

Results = <http://bio-aims/Results/17562517431e642d57/>

2013-04-21 20:37:27

PDBChain	LIBP Prediction
1QG HK	0.00%
1I4M*	29.66%
2QZTB	100.00%
1B0U*	45.34%

* the input contains no chain => LIBPpred used the entire protein (all the chains)

Done!

Figura 35: Ejemplo de utilización del servidor **LIBPpred**

3. CONCLUSIONES

Se exponen las conclusiones en concordancia con los objetivos trazados, agrupadas según el tipo de estudios realizados u objetivo perseguido: 1) desarrollo de programas, 2) búsqueda de modelos QSAR, 3) implementación de servidores, 4) publicación de resultados:

1. Se desarrollaron tres nuevas herramientas informáticas como programas de ordenador para el cálculo de índices topológicos de utilidad en el desarrollo de modelos QSAR a distintos niveles estructurales.
2. Se encontraron nuevos modelos QSAR aplicables a la predicción de la actividad biológica de compuestos de interés en Química Farmacéutica, Microbiología y Parasitología usando los nuevos programas desarrollados.
3. Se han implementado los nuevos modelos QSAR en cuatro herramientas informáticas para usar en la red (servidores “Web”), para la predicción “online” de la actividad biológica de compuestos y sus correspondientes dianas moleculares. Esto tiene un gran interés, sobre todo, en Química Farmacéutica, Microbiología y Parasitología.
4. Se publicaron los resultados en artículos de revistas especializadas y en capítulos de libro, describiendo las aplicaciones de las herramientas desarrolladas.
5. Se llevó a cabo la protección de la propiedad intelectual mediante los correspondientes registros de software.

Conclusión general:

Se puede concluir que las herramientas informáticas basadas en técnicas y procedimientos de ingeniería informática e inteligencia artificial, pueden ser de gran utilidad para el descubrimiento de fármacos y dianas moleculares.

4. REFERENCIAS

- [1] Gonzalez-Diaz H, Gonzalez-Diaz Y, Santana L, Ubeira FM, Uriarte E. Proteomics, networks and connectivity indices. *Proteomics* 2008; 8(4): 750-78.
- [2] Gonzalez-Diaz H. Quantitative studies on Structure-Activity and Structure-Property Relationships (QSAR/QSPR). *Curr Top Med Chem* 2008; 8(18): 1554.
- [3] 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.
- [4] Wang JF, Wei DQ, Chou KC. Pharmacogenomics and personalized use of drugs. *Curr Top Med Chem* 2008; 8(18): 1573-9.
- [5] 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.
- [6] 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.
- [7] Helguera AM, Combes RD, Gonzalez MP, Cordeiro MN. Applications of 2D descriptors in drug design: a DRAGON tale. *Curr Top Med Chem* 2008; 8(18): 1628-55.
- [8] Wang JF, Wei DQ, Chou KC. Drug candidates from traditional chinese medicines. *Curr Top Med Chem* 2008; 8(18): 1656-65.
- [9] 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.
- [10] 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.
- [11] Ivanciuc O. Weka machine learning for predicting the phospholipidosis inducing potential. *Curr Top Med Chem* 2008; 8(18): 1691-709.
- [12] Chen J, Shen B. Computational Analysis of Amino Acid Mutation: a Proteome Wide Perspective. *Curr Proteomics* 2009; 6(4): 228-34.
- [13] Chou KC. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr Proteomics* 2009; 6(4): 262-74.
- [14] Giuliani A, Di Paola L, Setola R. Proteins as Networks: A Mesoscopic Approach Using Haemoglobin Molecule as Case Study. *Curr Proteomics* 2009; 6(4): 235-45.
- [15] González-Díaz H, Prado-Prado F, Pérez-Montoto LG, Duardo-Sánchez A, López-Díaz A. QSAR Models for Proteins of Parasitic Organisms, Plants and Human Guests: Theory, Applications, Legal Protection, Taxes, and Regulatory Issues. *Curr Proteomics* 2009; 6(4): 214-27.
- [16] Ivanciuc O. Machine learning Quantitative Structure-Activity Relationships (QSAR) for peptides binding to Human Amphiphysin-1 SH3 domain. *Curr Proteomics* 2009; 6(4): 289-302.
- [17] Pérez-Montoto LG, Prado-Prado F, Ubeira FM, González-Díaz H. Study of Parasitic Infections, Cancer, and other Diseases with Mass-Spectrometry and Quantitative Proteome-Disease Relationships. *Curr Proteomics* 2009; 6(4): 246-61.
- [18] Torrens F, Castellano G. Topological Charge-Transfer Indices: From Small Molecules to Proteins. *Curr Proteomics* 2009; 6(4): 204-13.
- [19] Vázquez JM, Aguiar V, Seoane JA, Freire A, Serantes JA, Dorado J, et al. Star Graphs of Protein Sequences and Proteome Mass Spectra in Cancer Prediction. *Curr Proteomics* 2009; 6(4): 275-88.

- [20] Chou KC. Graphic rule for drug metabolism systems. *Curr Drug Metab* 2010; 11(4): 369-78.
- [21] Garcia I, Diop YF, Gomez G. QSAR & complex network study of the HMGR inhibitors structural diversity. *Curr Drug Metab* 2010; 11(4): 307-14.
- [22] Gonzalez-Diaz H. Network topological indices, drug metabolism, and distribution. *Curr Drug Metab* 2010; 11(4): 283-4.
- [23] Gonzalez-Diaz H, Duardo-Sanchez A, Ubeira FM, Prado-Prado F, Perez-Montoto LG, Concu R, 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(4): 379-406.
- [24] Khan MT. Predictions of the ADMET properties of candidate drug molecules utilizing different QSAR/QSPR modelling approaches. *Curr Drug Metab* 2010; 11(4): 285-95.
- [25] Martinez-Romero M, Vazquez-Naya JM, Rabunal JR, Pita-Fernandez S, Macenlle R, Castro-Alvarino J, et al. Artificial intelligence techniques for colorectal cancer drug metabolism: ontology and complex network. *Curr Drug Metab* 2010; 11(4): 347-68.
- [26] 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(4): 315-41.
- [27] Wang JF, Chou KC. Molecular modeling of cytochrome P450 and drug metabolism. *Curr Drug Metab* 2010; 11(4): 342-6.
- [28] 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(4): 296-306.
- [29] González-Díaz H. QSAR and Complex Networks in Pharmaceutical Design, Microbiology, Parasitology, Toxicology, Cancer, and Neurosciences. *Curr Pharm Des* 2010; 16(24): 2598-600.
- [30] 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.
- [31] Garcia I, Fall Y, Gomez G. QSAR, Docking, and CoMFA Studies of GSK3 Inhibitors. *Curr Pharm Des* 2010; 16(24): 2666-75.
- [32] 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(24): 2676-709.
- [33] 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 Des* 2010; 16(24): 2710-23.
- [34] Vazquez-Naya JM, Martinez-Romero M, Porto-Pazos AB, Novoa F, Valladares-Ayerbes M, Pereira J, et al. Ontologies of drug discovery and design for neurology, cardiology and oncology. *Curr Pharm Des* 2010; 16(24): 2724-36.
- [35] Gonzalez-Diaz H, Romaris F, Duardo-Sanchez A, Perez-Montoto LG, Prado-Prado F, Patlewicz G, et al. Predicting drugs and proteins in parasite infections with topological indices of complex networks: theoretical backgrounds, applications, and legal issues. *Curr Pharm Des* 2010; 16(24): 2737-64.
- [36] 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.
- [37] 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.
- [38] Munteanu CR, Fernandez-Blanco E, Seoane JA, Izquierdo-Novo P, Rodriguez-Fernandez JA, Prieto-Gonzalez JM, et al. Drug discovery and design for complex diseases through QSAR computational methods. *Curr Pharm Des* 2010; 16(24): 2640-55.

- [39] 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.
- [40] 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.
- [41] 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.
- [42] 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.
- [43] Bhattacharjee B, Jayadeepa RM, Banerjee S, Joshi J, Middha SK, Mole JP, et al. 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.
- [44] 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.
- [45] Wan SB, Hu LL, Niu S, Wang K, Cai YD, Lu WC, et al. Identification of multiple subcellular locations for proteins in budding yeast. *Current Bioinformatics* 2011; 6(1): 71-80.
- [46] 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.
- [47] Riera-Fernández P, Munteanu CR, Pedreira-Souto N, Martín-Romalde R, Duardo-Sanchez A, González-Díaz H. Definition of Markov-Harary Invariants and Review of Classic Topological Indices and Databases in Biology, Parasitology, Technology, and Social-Legal Networks. *Current Bioinformatics* 2011; 6(1): 94-121.
- [48] Dave K, Banerjee A. Bioinformatics analysis of functional relations between CNPs regions. *Current Bioinformatics* 2011; 6(1): 122-8.
- [49] Breiger R. The Analysis of Social Networks. In: *Handbook of Data Analysis*; Hardy M, Bryman A, eds., Sage Publications: London 2004; 505-26.
- [50] Abercrombie N, Hill S, Turner BS. Social structure. In: *The Penguin Dictionary of Sociology* 4th ed, Penguin: London 2000.
- [51] Craig C. Social Structure. In: *Dictionary of the Social Sciences*, Oxford University Press: Oxford 2002.
- [52] White H, Scott Boorman and Ronald Breiger. . ." Social Structure from Multiple Networks: I Blockmodels of Roles and Positions. *American Journal of Sociology* 1976; 81: 730-80.
- [53] Wellman B, Berkowitz SD. *Social Structures: A Network Approach*. Cambridge University Press: Cambridge 1988.
- [54] Newman MEJ. The structure and function of complex networks. *SIAM Review* 2003; 45: 167-256.
- [55] Bornholdt S, Schuster HG. *Handbook of Graphs and Complex Networks: From the Genome to the Internet*. WILEY-VCH GmbH & CO. KGa.: Wheinheim 2003.
- [56] Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. Wiley-VCH 2002.
- [57] Mauri A, Consonni V, Pavan M, Todeschini R. DRAGON Software: An Easy Approach to Molecular Descriptor Calculations. *MATCH, communications in mathematical and in computer chemistry* 2006; 56: 237-48.
- [58] Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, et al. Virtual computational chemistry laboratory – design and description. *J Comput Aided Mol Des* 2005; 19: 453–63.

- [59] Ponce YM. Total and local (atom and atom type) molecular quadratic indices: significance interpretation, comparison to other molecular descriptors, and QSPR/QSAR applications. *Bioorg Med Chem* 2004; 12(24): 6351-69.
- [60] Casanola-Martin GM, Marrero-Ponce Y, Khan MT, Ather A, Khan KM, Torrens F, et al. Dragon method for finding novel tyrosinase inhibitors: Biosilico identification and experimental in vitro assays. *Eur J Med Chem* 2007; 42(11-12): 1370-81.
- [61] Perez-Garrido A, Helguera AM, Rodriguez FG, Cordeiro MN. QSAR models to predict mutagenicity of acrylates, methacrylates and alpha,beta-unsaturated carbonyl compounds. *Dent Mater*; 26(5): 397-415.
- [62] Estrada E, Quincoces JA, Patlewicz G. Creating molecular diversity from antioxidants in Brazilian propolis. Combination of TOPS-MODE QSAR and virtual structure generation. *Mol Divers* 2004; 8(1): 21-33.
- [63] Cabrera-Pérez MA, Bermejo-Sanz M, Ramos-Torres L, Grau-Ávalos R, Pérez-González M, González-Díaz H. A topological sub-structural approach for predicting human intestinal absorption of drugs. *Eur J Med Chem* 2004; 39: 905-16.
- [64] Molina-Ruiz R, Saiz-Urra L, Rodriguez-Borges JE, Perez-Castillo Y, Gonzalez MP, Garcia-Mera X, et al. A TOPological Sub-structural Molecular Design (TOPS-MODE)-QSAR approach for modeling the antiproliferative activity against murine leukemia tumor cell line (L1210). *Bioorg Med Chem* 2009; 17(2): 537-47.
- [65] Casanola-Martin GM, Marrero-Ponce Y, Tareq Hassan Khan M, Torrens F, Perez-Gimenez F, Rescigno A. Atom- and bond-based 2D TOMOCOMD-CARDD approach and ligand-based virtual screening for the drug discovery of new tyrosinase inhibitors. *J Biomol Screen* 2008; 13(10): 1014-24.
- [66] Gonzalez-Diaz H, Duardo-Sanchez A, Ubeira FM, Prado-Prado F, Perez-Montoto LG, Concu R, 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.
- [67] Kier LB, Hall LH. *Molecular Structure Description: The Electrotopological State*. Academic Press 1999.
- [68] Katritzky AR, Oliferenko A, Lomaka A, Karelson M. Six-membered cyclic ureas as HIV-1 protease inhibitors: a QSAR study based on CODESSA PRO approach. Quantitative structure-activity relationships. *Bioorg Med Chem Lett* 2002; 12(23): 3453-7.
- [69] Katritzky AR, Kulshyn OV, Stoyanova-Slavova I, Dobchev DA, Kuanar M, Fara DC, et al. Antimalarial activity: a QSAR modeling using CODESSA PRO software. *Bioorg Med Chem* 2006; 14(7): 2333-57.
- [70] Katritzky AR, Dobchev DA, Tulp I, Karelson M, Carlson DA. QSAR study of mosquito repellents using Codessa Pro. *Bioorg Med Chem Lett* 2006; 16(8): 2306-11.
- [71] Katritzky AR, Kulshyn OV, Stoyanova-Slavova I, Dobchev DA, Kuanar M, Fara DC, et al. Antimalarial activity: a QSAR modeling using CODESSA PRO software. *Bioorganic & medicinal chemistry* 2006; 14(7): 2333-57.
- [72] Prado-Prado FJ, Borges F, Uriarte E, Perez-Montoto LG, Gonzalez-Diaz H. Multi-target spectral moment: QSAR for antiviral drugs vs. different viral species. *Anal Chim Acta* 2009; 651(2): 159-64.
- [73] Prado-Prado FJ, Martinez de la Vega O, 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(2): 569-75.
- [74] Prado-Prado FJ, Gonzalez-Diaz H, Santana L, Uriarte E. Unified QSAR approach to antimicrobials. Part 2: predicting activity against more than 90 different species in order to halt antibacterial resistance. *Bioorg Med Chem* 2007; 15(2): 897-902.

- [75] Prado-Prado FJ, Uriarte E, Borges F, Gonzalez-Diaz H. Multi-target spectral moments for QSAR and Complex Networks study of antibacterial drugs. *Eur J Med Chem* 2009; 44(11): 4516-21.
- [76] 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.
- [77] Prado-Prado FJ, Ubeira FM, Borges F, Gonzalez-Diaz H. Unified QSAR & network-based computational chemistry approach to antimicrobials. II. Multiple distance and triadic census analysis of antiparasitic drugs complex networks. *J Comput Chem* 2009.
- [78] 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.
- [79] Gonzalez-Diaz H, Prado-Prado FJ, Santana L, Uriarte E. Unify QSAR approach to antimicrobials. Part 1: predicting antifungal activity against different species. *Bioorg Med Chem* 2006; 14(17): 5973-80.
- [80] Gonzalez-Diaz H, Prado-Prado FJ. Unified QSAR and network-based computational chemistry approach to antimicrobials, part 1: multispecies activity models for antifungals. *J Comput Chem* 2008; 29(4): 656-67.
- [81] 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.
- [82] Shen HB, Chou KC. Virus-mPLoc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. *J Biomol Struct Dyn* 2010; 28(2): 175-86.
- [83] Shen HB, Chou KC. Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. *J Theor Biol* 2010; 264(2): 326-33.
- [84] Shen HB, Chou KC. Gpos-mPLoc: a top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins. *Protein Pept Lett* 2009; 16(12): 1478-84.
- [85] Shen HB, Chou KC. HIVcleave: a web-server for predicting HIV protease cleavage sites in proteins. *Anal Biochem* 2008; 375: 388-90.
- [86] Chou KC. Prediction of HIV protease cleavage sites in proteins. *Anal Biochem* 1996; 233: 1-14.

5. PUBLICACIONES (ANEXOS)

A continuación se presenta un ANEXO con las publicaciones que se recogen en la Tesis siguiendo el orden establecido en la misma.



Random Forest classification based on star graph topological indices for antioxidant proteins

Enrique Fernández-Blanco*, Vanessa Aguiar-Pulido, Cristian Robert Munteanu, Julian Dorado

University of A Coruña, ICT Dept., Facultad de Informática, Campus de Elviña s/n, 15071 A Coruña, Spain

HIGHLIGHTS

- ▶ This work presents an automatic antioxidant protein detection method.
- ▶ The new method uses graphical information processing theory which has never previously used in this kind of problem.
- ▶ The results can be qualified as notable compared with the state of the art.

ARTICLE INFO

Article history:

Received 9 July 2012

Received in revised form

17 September 2012

Accepted 2 October 2012

Available online 29 October 2012

Keywords:

Multi-target QSAR

Star Graph

Topological indices

Antioxidant protein

ABSTRACT

Aging and life quality is an important research topic nowadays in areas such as life sciences, chemistry, pharmacology, etc. People live longer, and, thus, they want to spend that extra time with a better quality of life. At this regard, there exists a tiny subset of molecules in nature, named antioxidant proteins that may influence the aging process. However, testing every single protein in order to identify its properties is quite expensive and inefficient. For this reason, this work proposes a model, in which the primary structure of the protein is represented using complex network graphs that can be used to reduce the number of proteins to be tested for antioxidant biological activity. The graph obtained as a representation will help us describe the complex system by using topological indices. More specifically, in this work, Randić's Star Networks have been used as well as the associated indices, calculated with the S2SNet tool. In order to simulate the existing proportion of antioxidant proteins in nature, a dataset containing 1999 proteins, of which 324 are antioxidant proteins, was created. Using this data as input, Star Graph Topological Indices were calculated with the S2SNet tool. These indices were then used as input to several classification techniques. Among the techniques utilised, the Random Forest has shown the best performance, achieving a score of 94% correctly classified instances. Although the target class (antioxidant proteins) represents a tiny subset inside the dataset, the proposed model is able to achieve a percentage of 81.8% correctly classified instances for this class, with a precision of 81.3%.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Life expectancy is increasing every year, especially in developed societies. Nowadays, in these countries, it is not strange to find some people that are near one hundred years, when 20 years ago this was quite rare. For example, in Spain, life expectancy at birth has increased from 73 years in 1975 to more than 81 in 2011 (OECD, 2011). In this context, it is obvious that people may want to spend the biggest part of their life in

optimum health conditions. In order to achieve this objective, finding some mechanism that delays aging (Cevenini et al., 2010; de Magalhães, 2010, 2011, 2012; Freitas and de Magalhães, 2012; Harman, 1981; Hayflick, 2000) is necessary. Several important works have proposed specific relationships between genes or proteins and aging (Aledo et al., 2011, 2012; de Magalhães et al., 2009; Freitas et al., 2011; Gomes et al., 2011; Li et al., 2010).

More research focused on antioxidant molecules may be useful for this purpose, since, for example, oxidative stress is one of the risk factors of colorectal carcinogenesis. In inflammatory reactions the activated leucocytes produce mutagenic and mitogenic free radicals, hereby promoting tumour formation. In addition, obesity, hyperlipidemia and hyperinsulinemia increase the energy supply of epithelial cells, thus leading to deregulation of the mitochondrial electron transport chain. Finally, the latter

* Corresponding author at: University of A Coruña, ICT Dept., Facultad de Informática, Campus de Elviña s/n, 15071 A Coruña, Spain. Tel.: +34 981 167 000; fax: +34 981 167 160.

E-mail addresses: efernandez@udc.es (E. Fernández-Blanco), vaguiar@udc.es (V. Aguiar-Pulido), muntisa@gmail.com (C.R. Munteanu), julian@udc.es (J. Dorado).

leads to increased free radical production, causing troubles in cell cycle regulation, mutations, and unrestricted proliferation of damaged cells (Regöly-Mérei et al., 2007).

Unfortunately, the number of molecules that have antioxidant properties in nature is quite low. Therefore, developing models that help to detect molecules with antioxidant properties would be very helpful. On this basis, the main objective of this paper will be to develop models that, on one hand, will reduce the number of molecules for tests in different trials and, on the other hand, to increase the success rates when molecules are tested looking for these properties.

In order to achieve this, the authors have used Quantitative Structure Activity Relationships (QSARs) (Devillers and Balaban, 1999). QSARs are based on Graph Theory, one of the most common techniques used in protein analysis. Using this technique, macromolecular descriptors, named topological indexes (TIs), are calculated for its later analysis. This branch of mathematical chemistry has become an intense area of research, generating new information regarding DNA/proteins by representing them as graphs and obtaining the corresponding TIs in order to analyse the resulting complex networks (Agüero-Chapin et al., 2006; Bielińska-Wa-z et al., 2007; Munteanu et al., 2010; Randić and Balaban, 2003). In order to perform these analyses, the TIs are then processed by a classification technique such as Support Vector Machines (SVMs) (Vapnik, 1995), Artificial Neural Networks (ANNs) (Rivero et al., 2011), Random Space Classifiers (Skurichina and Duin, 2002), Linear Discriminant Analysis (LDA), etc, abstracting general properties for future molecules that have not been already tested. Many examples involving QSAR can be found in literature (González-Díaz et al., 2006, 2007a, 2010; Prado-Prado et al., 2008; Riera-Fernández et al., 2012) regarding protein folding kinetics (Chou, 1990), enzyme-catalyzed reactions (Chou, 1989; Chou and Forsen, 1980; Chou and Liu, 1981; Kuzmic et al., 1992), inhibition kinetics of processive nucleic acid polymerases and nucleases (Althaus et al., 1993a, 1993b, 1994, 1996; Chou et al., 1994), DNA sequence analysis (Qi et al., 2007), anti-sense strands base frequencies (Chou et al., 1996), analysis of codon usage (Chou and Zhang, 1992; Zhang and Chou, 1994), Cancer prediction (Aguiar-Pulido et al., 2012), as well as complex network systems investigations (Diao et al., 2007; Gonzalez-Diaz et al., 2007b, 2008).

In this work, the authors propose the first non-antioxidant/antioxidant protein classification model based on embedded/ non-embedded Star Graph TIs including the trace of connectivity matrices, Harary number, Wiener index, Gutman index, Schultz index, Moreau-Broto indices, Balaban distance connectivity index, Kier–Hall connectivity indices and Randić connectivity index. This information is then used as input to several classification techniques, obtaining the best results when the Random Forest technique is used.

2. Materials and methods

The description of the methodology followed in this work is presented in Fig. 1. The input data is represented by the amino acid sequences (primary structure) antioxidant and non-antioxidant proteins in FASTA format. By using the S2SNet tool (Munteanu et al., 2009), the sequences of amino acids are transformed into Star Graphs and the corresponding topological indices are calculated. The resulting numbers that characterised each graph (that is, a protein graphical representation) are then used in Weka (Hall et al., 2009a) to find the best QSAR classification model. The final model is used to predict antioxidant activity for new amino acid sequences.

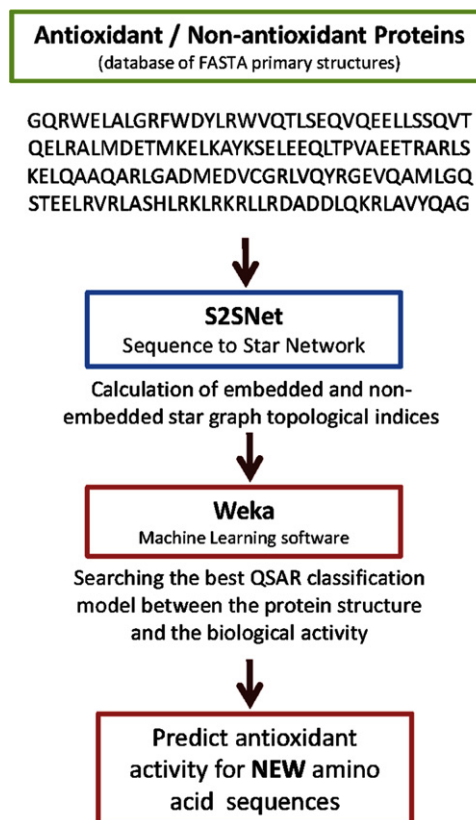


Fig. 1. Flowchart of building QSAR classification models for protein antioxidant activity prediction.

2.1. Protein set

This work is based on datasets extracted from several protein databases. The sets of protein primary sequences are represented by 324 proteins with antioxidant activity and 1675 proteins without. The antioxidant protein FASTA sequences (positive group) have been downloaded from the Protein Databank (Berman et al., 2000), the “Antioxidant activity” list obtained with the “Molecular Function Browser” in the “Advanced Search Interface”. The negative group was constructed using the PISCES CullerPDB (Wang and Dunbrack, 2003) list of proteins with identity less than 20%, resolution of 1.6 Å and *R*-factor 0.25 (non-antioxidant proteins included, but any other possible biological function). Identity is the degree of correspondence between two sequences and a value of 25% or higher implies similarity of function. The sequence identities for PDB sequences have been determined using Combinatorial Extension (CE) structural alignment (Shindyalov and Bourne, 1998). The PIECES server (<http://dunbrack.fccc.edu/PISCES.php>) used a Z-score of 3.5 as the threshold to accept possible evolutionary relationships. PISCES’ alignments are local, so that two proteins that share a common domain with sequence identity above the threshold are not both included in the output lists. Both lists have not been post-filtered for any source organism.

2.2. Star Graph topological indices

Each protein was transformed into a Star Graph, where the amino acids are the vertices (nodes), connected in a specific sequence by the peptide bonds. The Star Graph is a special type of tree with *N* vertices where one has got *N*-1 degrees of freedom and the remaining *N*-1 vertices have got one single degree of freedom (Harary, 1969). Each of the 20 possible branches (“rays”)

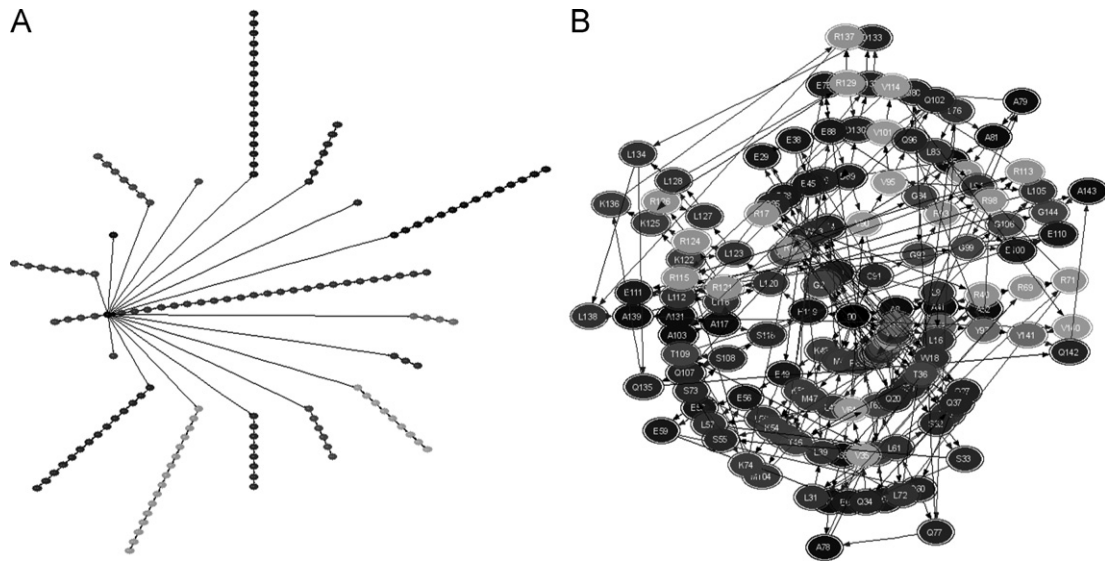


Fig. 2. The non-embedded (A) and embedded (B) Star Graphs for 1BZ4, chain A.

of the star contains the same amino acid type and the star centre is a non-amino acid vertex. This way, the following information of the protein primary structure is encoded into the Star Graph connectivity: amino acid type, sequence and frequency.

A protein can be represented by diverse forms of graphs, which can be associated with distinct distance matrices. The best method to construct a standard Star Graph is described subsequently: each amino acid/vertex holds the position in the original sequence and the branches are labelled by alphabetical order of the three-letter amino acid code (Randić et al., 2007). The graph is embedded if the initial sequence connectivity in the protein chain is included. Fig. 2 presents the embedded/ non-embedded Star Graphs of PRPS1 using the alphabetical order of one-letter amino acid code.

Graphs are compared using the corresponding connectivity matrix, distance matrix and degree matrix. In the case of the embedded graph, the connectivity matrices in the sequence and in the Star Graph are combined. These matrices and the normalized ones are the basis of the TIs calculation.

The conversion of the amino acid sequences into Star Graph TIs was performed by using the Sequence to the Star Networks (S2SNet) application, developed by our group. S2SNet is based on wxPython (Rappin and Dunn, 2006) for the GUI application and has Graphviz (Koutsofios and North, 1993) as a graphics back-end. The present calculations are characterized by embedded and non-embedded TIs, no weights, Markov normalization and power of matrices/indices (n) up to 5. The results file contains the following TIs (Todeschini and Consonni, 2002):

Trace of the n connectivity matrices (Tr_n):

$$Tr_n = \sum_i (M^n)_{ii}, \quad (1)$$

where $n=0$ – power limit, M =graph connectivity matrix (i^*i dimension); ii = i th diagonal element;

Harary number (H):

$$H = \sum_{i < j} m_{ij}/d_{ij}, \quad (2)$$

where d_{ij} are the elements of the distance matrix and m_{ij} are the elements of the M connectivity matrix;

Wiener index (W):

$$W = \sum_{i < j} d_{ij}, \quad (3)$$

Gutman topological index (S_6):

$$S_6 = \sum_{ij} \text{deg}_i \times \text{deg}_j / d_{ij}, \quad (4)$$

where deg_i are the elements of the degree matrix;
Schultz topological index (non-trivial part) (S):

$$S = \sum_{i < j} (\text{deg}_i + \text{deg}_j) \times d_{ij}, \quad (5)$$

Balaban distance connectivity index (J):

$$J = (\text{edges} - \text{nodes} + 2) \times \sum_{i < j} m_{ij} \times \sqrt{\sum_k d_{ik} \times \sum_k d_{kj}}, \quad (6)$$

where $\text{nodes} + 1 = \text{AA numbers/node number in the Star Graph} + \text{origin}$, $\sum_k d_{ik}$ is the node distance degree;

Kier–Hall connectivity indices (nX):

$$^0X = \sum_i 1/\sqrt{\text{deg}_i}, \quad (7)$$

$$^2X = \sum_{i < j < k} m_{ij} \times m_{jk} / \sqrt{(\text{deg}_i \times \text{deg}_j \times \text{deg}_k)}, \quad (8)$$

$$^3X = \sum_{i < j < k < m} m_{ij} \times m_{jk} \times m_{km} / \sqrt{(\text{deg}_i \times \text{deg}_j \times \text{deg}_k \times \text{deg}_m)}, \quad (9)$$

$$^4X = \sum_{i < j < k < m < o} m_{ij} \times m_{jk} \times m_{km} \times m_{mo} / \sqrt{(\text{deg}_i \times \text{deg}_j \times \text{deg}_k \times \text{deg}_m \times \text{deg}_o)}, \quad (10)$$

$$^5X = \sum_{i < j < k < m < o < q} m_{ij} \times m_{jk} \times m_{km} \times m_{mo} \times m_{oq} / \sqrt{(\text{deg}_i \times \text{deg}_j \times \text{deg}_k \times \text{deg}_m \times \text{deg}_o \times \text{deg}_q)}, \quad (11)$$

Randić connectivity index (1X):

$$^1X = \sum_{ij} m_{ij} / \sqrt{(\text{deg}_i \times \text{deg}_j)}, \quad (12)$$

These TIs and other derivate ones will be used in the next step to construct an antioxidant/ non-antioxidant classification model using machine learning methods.

2.3. Random Forest

Random Forest was first proposed by Breiman, (2001). This technique combines many decision trees to make a prediction, giving as output the class that is the mode of the classes output by

individual trees. Thus, this technique can be considered an “ensemble learning” technique, since it uses multiple models to obtain better predictive performance. These decision trees are constructed by means of bagging classification trees (Breiman, 1996), where each tree is constructed independently based on a random sample and a majority vote of the trees is taken as prediction. Random Forest adds an extra random layer to bagging. Normally, decision trees are built from a random sample and nodes are split by the best among a subset of predictors randomly chosen at that node.

The main advantage of Random Forest over other techniques such as Artificial Neural Networks, Support Vector Machines, Linear Discriminant Analysis, etc. is the robustness of this technique regarding solution overfitting, tending to converge always when the number of trees is large.

The typical Random Forest algorithm is composed of three steps:

- Get n random samples from the original dataset to use them as tree seeds.
- For each seed, grow a non-pruned tree, and for each node, randomly choose m predictors and the best split among those.
- Execute the different prediction trees and select as prediction the most voted one.

It may be highlighted that this technique is quite efficient because, when constructing the trees, the pruning phase has been deleted and the search is performed over a small set. This simplification can give the idea that a single tree may have better performance, but it was empirically proved that Random Forest overcomes the performance of CART single tree predictors (Chipman et al., 1998).

3. Results

The dataset used in this paper is composed of 1999 protein sequences, from which 324 have proved to have antioxidant activity (positive group). The remaining 1675 proteins (negative group) are sequences from the CullerPDB server with identity less than 20%, without antioxidant biological activity. These protein sequences have been processed with the S2SNet application (Munteanu et al., 2009) in order to obtain the different topological indexes used in this study. Specifically, from each sequence 42 attributes are extracted from the embedded/non-embedded Star Graph.

The series of topological indices for each protein have been used to find the best antioxidant classification model with Machine Learning methods included in Weka (Hall et al., 2009b). In order to extract more general conclusions from this study, the authors have tested the different classification techniques using 10-fold cross-validation (McLachlan et al., 2004).

Table 1
Performance of the classification methods considering all the attributes.

Technique	% Antiox	% Non antiox	% Global	Precision antiox (%)	Global precision (%)	ROC
Naive bayes	97.5	49.1	57.0	27.1	87.4	0.78
MLP	22.8	97.5	85.4	63.8	83.0	0.874
K-star	86.7	94.3	93.1	74.7	93.7	0.971
JRip	64.8	96.1	91.0	76.1	90.6	0.814
Random tree	81.8	95.0	92.8	75.9	93.1	0.884
Random Forest	84	96.7	94.6	82.9	94.6	0.954

10-fold cross-validation is the most common among the k -fold cross-validation family and its objective is to minimize the influence of the randomness in creating the training and test sets for a specific classification technique.

The objective of this work is to select the technique with the highest classification score, having a good precision value, due to the nature of the problem. The first approach considered was to use linear regression, but the results showed that it was impossible to achieve good classification scores with this technique.

Table 1 shows the results of the different classification models tested, those that obtained the best scores, considering all the attributes extracted from the Star Graph, that is, 42 attributes. The algorithms used in the tests are those implemented in the Weka Machine Learning framework. This table shows, for each model, the classification scores obtained for the different classes, as well as the global classification percentages, the precision values for the target class (antioxidant proteins), the ROC values and the number of attributes that were considered.

The Random Forest technique seems to be the best option because it achieves a percentage of 94.6% correctly classified instances. In addition, it is interesting to note that, for the antioxidant class, it achieves a percentage of 84% correctly classified instances. This model achieves a precision of 82.9%, which is the highest among the tested machine learning methods.

Table 2
Attributes subsets for the tests.

Subset Name	Attributes	
	Non-embedded graph	Embedded graph
Sh	Sh0,Sh1, Sh2, Sh3, Sh4, Sh5	eSh0,eSh1, eSh2, eSh3, eSh4, eSh5
Tr	Tr0, Tr2, Tr4	eTr0, eTr2, eTr3,eTr4,eTr5
X	X0, X1R, X2, X3, X4, X5	eX0, eX1R, eX2, eX3, eX4, eX5
Remaining	H, W, S6, S, J	eH, eW, eS6, eS, eJ

Table 3
Results obtained using the different subsets as input, considering 12 attributes.

Technique	% antiox	% non antiox	% global	Precision antiOx (%)	Global precision (%)	ROC
Naive bayes	95.7	56.3	62.7	29.8	87.4	0.79
MLP	38.6	95.5	86.2	62.2	84.6	0.851
K-star	51.5	95.2	88.1	67.3	87.2	0.926
JRip	47.2	98.6	90.0	86.4	89.9	0.726
Random tree	80.9	94.2	92.0	73.0	92.5	0.875
Random Forest	79.3	94.4	91.9	73.2	92.3	0.913
Naive bayes	74.0	57.3	60.1	74.7	60.1	0.797
MLP	0	100	83.8	0	83.8	0.644
K-star	82.1	94.0	92.0	72.5	92.6	0.961
JRip	63.9	97.0	91.6	80.2	91.2	0.815
Random tree	79.0	94.3	91.8	72.9	92.2	0.867
Random Forest	79.9	96.1	93.5	79.9	93.5	0.95
Naive bayes	77.5	55.8	59.3	25.3	81.8	0.772
MLP	0	100	83.8	0	83.8	0.644
K-star	77.2	94.2	90.6	70.7	90.7	0.946
JRip	67.0	96.7	91.9	79.8	91.5	0.840
Random tree	82.1	94.9	92.8	75.6	93.1	0.885
Random Forest	82.1	96.1	93.8	80.4	93.9	0.948

In order to reduce the noise and to improve the classification scores, the data used as input has been divided into three subsets depending on the nature of the attributes:

- A subset named *Sh*, which includes the attributes related with the entropy of the embedded and non-embedded Graph.
- A subset named *Tr*, which includes the attributes related with the traces of the embedded and non-embedded Graph.
- And a subset named *X*, which includes the attributes related with the polygon indexes to represent the subspaces in the graph.

Table 2 shows the result of this division. It should be highlighted that not all of the original attributes have been included in one of these three subsets; more specifically, some attributes regarding the general shape of the graphs were not included in any of these subsets.

The different methods were then tested using each of these subsets as well as their combination, in order to find the best possible one. Results of these tests are shown in Tables 3 and 4.

These results show that Random Forest can still be considered adequate to solve the problem proposed in this work and that there is nearly no difference between using the *X* subset as input and all of the attributes. Regarding classification scores, this technique achieves 82.1% of correctly classified instances for the target class (that is, the antioxidant class) with a precision of 80.4% considering the 12 attributes part of the *X* subset, compared to 84% of correctly classified instances with a precision of 82.9% when all the attributes were considered (that is, 42 attributes). Therefore, it is very likely that some of these attributes may give little extra information. Reducing the number of attributes considered as input may be interesting, improving even the performance or precision of the model.

After analysing the results shown above, it seems that Random Forest is the best and most robust classification model. As it was previously mentioned, the subsets *Sh*, *Tr* and *X* contain the properties of the embedded and non-embedded graph. Therefore, in order to try to reduce the number of input attributes, the authors have tested the Random Forest in more depth, distinguishing between the properties of both types of graph. Results regarding this are shown in Table 5, as well as the number of attributes used as input to the method.

Table 4
Results obtained using combinations of the different subsets as input, considering 20 attributes.

Technique	% antiox	% non antiox	% global	Precision antiox (%)	Global precision (%)	ROC
Naive bayes	96.0	57.0	63.3	30.1	87.5	0.807
MLP	16.4	98.2	84.9	63.9	82.3	0.867
K-star	84.3	93.9	92.3	72.8	93.0	0.967
JRip	65.7	97.0	91.9	81.0	91.6	0.843
Random tree	82.4	94.9	92.8	75.6	93.1	0.886
Random Forest	81.8	96.5	94.1	81.8	94.1	0.947
Naive bayes	80.6	55.5	59.5	25.9	82.7	0.783
MLP	38.9	95.2	86.1	61.2	84.5	0.877
K-star	78.4	94.4	91.8	73.2	92.1	0.957
JRip	65.1	96.8	91.7	79.9	91.3	0.836
Random tree	81.8	95.3	93.1	77.3	93.3	0.886
Random forest	81.2	95.7	93.3	78.5	93.4	0.952
Naive bayes	78.4	54.2	58.1	24.9	81.8	0.792
MLP	0	100	83.8	0	83.8	0.644
K-star	86.4	93.7	92.5	72.5	93.3	0.97
JRip	68.2	96.5	91.9	78.9	91.6	0.846
Random tree	81.8	94.7	92.6	74.9	92.6	0.882
Random forest	83.6	96.9	94.7	83.9	94.7	0.951

Table 5
Scores obtained by the Random Forest method for each input dataset tested.

Subset	% antiox	% non antiox	% global	Precision antiox (%)	Global precision (%)	ROC	Number attributes
Sh	79.3	94.4	91.9	73.2	92.3	0.913	12
Sh-embedded	79.0	94.1	91.6	72.1	92	0.897	6
Sh-non-embedded	75.0	94.6	91.4	73.0	91.5	0.906	6
Tr	79.9	96.1	93.5	79.9	93.5	0.95	8
Tr-embedded	81.8	96.4	94.0	81.3	94.0	0.954	5
TR-non-embedded	79.9	94.0	91.7	72.1	92.2	0.903	3
X	82.1	96.1	93.8	80.4	93.9	0.948	12
X-embedded	82.4	95.7	92.5	78.8	93.7	0.938	6
X-non-embedded	79.9	95.2	92.7	76.2	92.9	0.926	6
Sh and Tr	81.8	96.5	94.1	81.8	94.1	0.947	20
Sh- and Tr-embedded	81.2	96.0	93.6	79.7	93.6	0.946	11
Sh- and Tr-non-embedded	79.6	95.5	92.9	77.5	93.0	0.927	9
Sh and X	81.2	95.7	93.3	78.5	93.4	0.952	24
Sh- and X-embedded	80.2	95.1	92.7	76.0	92.9	0.947	12
Sh- and X-non-embedded	79.6	95.5	92.9	77.5	93.0	0.927	12
Tr and X	83.6	96.9	94.7	83.9	94.7	0.951	20
Tr- and X-embedded	83.6	96.8	94.6	83.6	94.7	0.958	11
Tr- and X-non-embedded	80.2	95.5	93.0	77.4	93.1	0.935	9
All	84	96.7	94.6	82.9	94.6	0.954	42
All-embedded	82.1	96.8	94.4	83.1	94.4	0.954	22
All-non-embedded	81.2	95.6	93.2	78.0	93.4	0.934	20

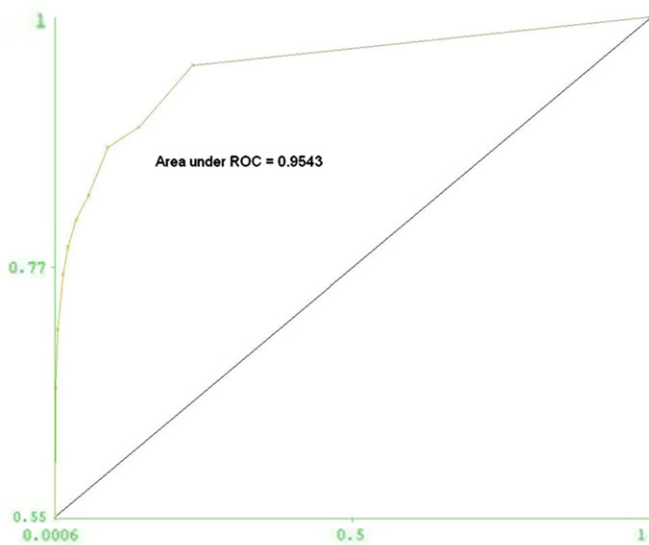


Fig. 3. ROC curve plot for the best classification method and the dataset containing the smallest number of attributes.

Again, results show that Random Forest is able to achieve better classification scores and similar precision values considering less attributes as input; in this case, taking only into consideration those included in the *Tr* subset (which contains only the values of the embedded graph). By adding the embedded attributes of the *X* subset, results are somehow better. However, this implies doubling the number of attributes used as input to the model. Thus, these results confirm that the rest of the attributes seem to add very little information or may even introduce noise inducing worse classification scores. If the ROC value is checked, it can be observed that the same ROC values are obtained when using the *Tr*-embedded dataset and the dataset containing all the attributes. The ROC curve for the *Tr*-embedded dataset is shown in Fig. 3.

4. Discussion

This study proposes a model designed to identify proteins that have antioxidant activity by using Star Graph TIs obtained from protein amino acid sequences. The proposed model, based on only five attributes extracted from the embedded graph, shows good predictive capacity, achieving 94% of correctly classified instances. It is also important to highlight that, even though the non-antioxidant class was not the target class of this study, the model achieves a score of 81.8% correctly classified instances with good precision (81.3%).

Antioxidant proteins are very important molecules in pharmacology today. It can be concluded from this study that this model may help reducing the number of proteins to be tested in antioxidant research, being very probable that the selected proteins have antioxidant properties.

Acknowledgements

Vanessa Aguiar-Pulido and Cristian R. Munteanu acknowledge the funding support for a research position by the “Plan I2C” and an “Isidro Parga Pondal” Program both from Xunta de Galicia, Spain (supported by the European Social Fund). The authors also want to thank the support from different projects that has funded part of this research (CN 2011/034, CN2012/127, 10SIN105004PR, 09SIN010105PR and TIN-2009-07707).

References

- Agüero-Chapin, G., Gonzalez-Diaz, H., Molina, R., Varona-Santos, J., Uriarte, E., Gonzalez-Diaz, Y., 2006. Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from *Psidium guajava* L. *FEBS Lett.* 580, 723–730.
- Aguiar-Pulido, V., Munteanu, C.R., Seoane, J.A., Fernández-Blanco, E., Pérez-Montoto, L.G., González-Díaz, H., Dorado, J., 2012. Naïve Bayes QSDR classification based on spiral-graph Shannon entropies for protein biomarkers in human colon cancer. *Mol. Biosyst.* 8, 1716–1722.
- Aledo, J.C., Li, Y., de Magalhães, J.P., Ruiz-Camacho, M., Perez-Claros, J.A., 2011. Mitochondrially encoded methionine is inversely related to longevity in mammals. *Aging Cell* 10, 198–207.
- Aledo, J.C., Valverde, H., de Magalhães, J.P., 2012. Mutational bias plays an important role in shaping longevity-related amino acid content in Mammalian mtDNA-encoded proteins. *J. Mol. Evol.* 74, 332–341.
- Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1993a. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. *Biochemistry* 32, 6548–6554.
- Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1993b. Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. *J. Biol. Chem.* 268, 6119–6124.
- Althaus, I.W., Chou, J.J., Gonzales, A.J., LeMay, R.J., Deibel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Thomas, R.C., Aristoff, P.A., et al., 1994. Steady-state kinetic studies with the polysulfonate U-9843, an HIV reverse transcriptase inhibitor. *Experientia* 50, 23–28.
- Althaus, I.W., Chou, K.C., Lemay, R.J., Franks, K.M., Deibel, M.R., Kezdy, F.J., Resnick, L., Busso, M.E., So, A.G., Downey, K.M., Romero, D.L., Thomas, R.C., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1996. The benzylthio-pyrimidine U-31,355, a potent inhibitor of HIV-1 reverse transcriptase. *Biochem. Pharmacol.* 51, 743–750.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242.
- Bielińska-Wa-z, D., Nowak, W., Wa-z, P., Nandyc, A., Clark, T., 2007. Distribution-moments of 2D-graphs as descriptors of DNAsequences. *Chem. Phys. Lett.* 443, 408–413.
- Breiman, L., 1996. Bagging predictors. *Mach. Learn.* 24, 123–140.
- Breiman, L., 2001. Random Forest. *Mach. Learn.* 45, 5–32.
- Cevenini, E., Bellavista, E., Tieri, P., Castellani, G., Lescai, F., Francesconi, M., Mishto, M., Santoro, A., Valensin, S., Salvioli, S., Capri, M., Zaikin, A., Monti, D., de Magalhães, J.P., Franceschi, C., 2010. Systems biology and longevity: an emerging approach to identify innovative anti-aging targets and strategies. *Curr. Pharm. Des.* 16, 802–813.
- Chipman, H.A., George, E.I., McCulloch, R.E., 1998. An introduction to Classification and Regression Tree (CART) analysis. *J. Am. Stat. Assoc.* 935–948.
- Chou, K.C., 1989. Graphical rules in steady and non-steady enzyme kinetics. *J. Biol. Chem.* 264, 12074–12079.
- Chou, K.C., 1990. Review: applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. *Biophys. Chem.* 35, 1–24.
- Chou, K.C., Forsen, S., 1980. Graphical rules for enzyme-catalyzed rate laws. *Biochem. J.* 187, 829–835.
- 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., Liu, W.M., 1981. Graphical rules for non-steady state enzyme kinetics. *J. Theor. Biol.* 91, 637–654.
- Chou, K.C., Zhang, C.T., 1992. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. *AIDS Res. Hum. Retroviruses* 8, 1967–1976.
- Chou, K.C., Zhang, C.T., Elrod, D.W., 1996. Do “antisense proteins” exist? *J. Protein Chem.* 15, 59–61.
- de Magalhães, J.P., Curado, J., Church, G.M., 2009. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics* 25, 875–881.
- de Magalhães, J.P., Finch, C.E., Janssens, G., 2010. Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions. *Ageing Res. Rev.* 9, 315–323.
- de Magalhães, J.P., 2011. The biology of ageing: a primer. In: I., S.-H. (Ed.), *An Introduction to Gerontology*. Cambridge University Press, Cambridge, UK, pp. 21–47.
- de Magalhães, J.P., Wuttke, D., Wood, S.H., Plank, M., Vora, C., 2012. Genome-environment interactions that modulate aging: powerful targets for drug discovery. *Pharmacol. Rev.* 64, 88–101.
- Devillers, J., Balaban, A.T., 1999. Topological Indices and Related Descriptors in QSAR and QSPR. Gordon and Breach, The Netherlands.
- Diao, Y., Li, M., Feng, Z., Yin, J., Pan, Y., 2007. The community structure of human cellular signaling network. *J. Theor. Biol.* 247, 608–615.
- Freitas, A.A., de Magalhães, J.P., 2012. A review and appraisal of the DNA damage theory of ageing. *Mutat. Res.* 728, 12–22.
- Freitas, A.A., Vasieva, O., de Magalhães, J.P., 2011. A data mining approach for classifying DNA repair genes into ageing-related or non-ageing-related. *BMC Genomics.* 12, 27.

- Gomes, N.M., Ryder, O.A., Houck, M.L., Charter, S.J., Walker, W., Forsyth, N.R., Austad, S.N., Venditti, C., Pagel, M., Shay, J.W., Wright, W.E., 2011. Comparative biology of mammalian telomeres: hypotheses on ancestral states and the roles of telomeres in longevity determination. *Aging Cell* 10, 761–768.
- González-Díaz, H., Bonet, I., Terán, C., de Clercq, E., Bello, R., García, M., Santana, L., Uriarte, E., 2007a. ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. *Eur. J. Med. Chem.* 42, 580–585.
- González-Díaz, H., González-Díaz, Y., Santana, L., Ubeira, F.M., Uriarte, E., 2008. Proteomics, networks and connectivity indices. *Proteomics* 8, 750–778.
- González-Díaz, H., Sanchez-Gonzalez, A., Gonzalez-Diaz, Y., 2006. 3D-QSAR study for DNA cleavage proteins with a potential anti-tumor ATCUN-like motif. *J. Inorg. Biochem.* 100, 1290–1297.
- González-Díaz, H., Vilar, S., Rivero, D., Fernández-Blanco, E., Porto, A., Munteanu, C.R., 2010. QSPR Models for Cerebral Cortex Co-Activation Networks, Topological Indices for Medicinal Chemistry, Biology, Parasitology, and Social Networks. Research Signpost.
- González-Díaz, H., Vilar, S., Santana, L., Uriarte, E., 2007b. Medicinal chemistry and bioinformatics—current trends in drugs discovery with networks topological indices. *Curr. Top Med. Chem.* 7, 1025–1039.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.A., 2009a. The WEKA data mining software: an update. *SIGKDD Explor.*, 11.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009b. The WEKA data mining software: an update. *SIGKDD Explor.*, 11.
- Harary, F., 1969. *Graph Theory*. Reading, MA.
- Harman, D., 1981. The aging process. *Proc. Natl. Acad. Sci. U.S.A.* 78, 7124–7128.
- Hayflick, L., 2000. The future of ageing. *Nature* 408, 267–269.
- Koutsofios, E., North, S.C., 1993. *Drawing Graphs with Dot*. AT&T Bell Laboratories, Murray Hill, NJ, USA.
- Kuzmic, P., Ng, K.Y., Heath, T.D., 1992. Mixtures of tight-binding enzyme inhibitors. Kinetic analysis by a recursive rate equation. *Anal. Biochem.* 200, 68–73.
- Li, Y.H., Dong, M.Q., Guo, Z., 2010. Systematic analysis and prediction of longevity genes in *Caenorhabditis elegans*. *Mech. Ageing Dev.* 131, 700–709.
- McLachlan, G.J., Do, K.-A., Ambroise, C., 2004. *Analyzing Microarray Gene Expression Data*. Wiley.
- 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., 2010. Drug discovery and design for complex diseases through QSAR computational methods. *Curr. Pharm. Design* 16, 2640–2655.
- Munteanu, C.R., Magalhães, A.L., Uriarte, E., González-Díaz, H., 2009. Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. *J. Theor. Biol.* 257, 303–311.
- OECD, 2011. <http://stats.oecd.org/index.aspx?DataSetCode=HEALTH_STAT>.
- Prado-Prado, F.J., González-Díaz, H., Martínez de la Vega, O., Ubeira, F.M., Chou, K.C., 2008. 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.* 16, 5871–5880.
- Qi, X.Q., Wen, J., Qi, Z.H., 2007. New 3D graphical representation of DNA sequence based on dual nucleotides. *J. Theor. Biol.* 249, 681–690.
- Randić, M., Balaban, A.T., 2003. On a four-dimensional representation of DNA primary sequences. *J. Chem. Inf. Model.* 43, 532–539.
- Randić, M., Zupan, J., Vikić-Topić, D., 2007. On representation of proteins by star-like graphs. *J. Mol. Graph. Model.* 290–305.
- Rappin, N., Dunn, R., 2006. *wxPython in Action*. Manning Publications Co., Greenwich, CT.
- Regöly-Mérei, A., Bereczky, M., Arató, G., Telek, G., Pallai, Z., Lugasi, A., Antal, M., 2007. Nutritional and antioxidant status of colorectal cancer patients. *Orv. Hetil.* 148, 1505–1509.
- Riera-Fernández, I., Martín-Romalde, R., Prado-Prado, F., Escobar, M., Munteanu, C., Concu, R., Duardo-Sanchez, A., González-Díaz, H., 2012. From QSAR models of drugs to complex networks: state-of-art review and introduction of new Markov-spectral moments indices. *Curr. Top. Med. Chem.* 8, 927–960.
- Rivero, D., Fernandez-Blanco, E., Dorado, J., Pazos, A., 2011. Using recurrent ANNs for the detection of epileptic seizures in EEG signals. *Evolutionary Computation (CEC)*, 2011 IEEE Congress on IEEE, pp. 587–592.
- Shindyalov, I.N., Bourne, P.E., 1998. Protein structure alignment by incremental combinatorial extension of the optimum path. *Protein Eng.* 11, 739–747.
- Skurichina, M., Duin, R.P.W., 2002. Bagging, boosting and the random subspace method for linear classifiers. *Pattern Anal. Appl.* 5, 121–135.
- Todeschini, R., Consonni, V., 2002. *Handbook of Molecular Descriptors*. Wiley-VCH.
- Vapnik, V.N., 1995. *The Nature of Statistical Learning Theory*.
- Wang, G., Dunbrack Jr., R.L., 2003. PISCES: a protein sequence culling server. *Bioinformatics* 19, 1589–1591.
- Zhang, C.T., Chou, K.C., 1994. Analysis of codon usage in 1562 *E. coli* protein coding sequences. *J. Mol. Biol.* 238, 1–8.

Naïve Bayes QSDR classification based on spiral-graph Shannon entropies for protein biomarkers in human colon cancer

Vanessa Aguiar-Pulido,^a Cristian R. Munteanu,^a José A. Seoane,^{*a}
Enrique Fernández-Blanco,^a Lázaro G. Pérez-Montoto,^b Humberto González-Díaz^b
and Julián Dorado^a

Received 2nd February 2012, Accepted 9th March 2012

DOI: 10.1039/c2mb25039j

Fast cancer diagnosis represents a real necessity in applied medicine due to the importance of this disease. Thus, theoretical models can help as prediction tools. Graph theory representation is one option because it permits us to numerically describe any real system such as the protein macromolecules by transforming real properties into molecular graph topological indices. This study proposes a new classification model for proteins linked with human colon cancer by using spiral graph topological indices of protein amino acid sequences. The best quantitative structure–disease relationship model is based on eleven Shannon entropy indices. It was obtained with the Naïve Bayes method and shows excellent predictive ability (90.92%) for new proteins linked with this type of cancer. The statistical analysis confirms that this model allows diagnosing the absence of human colon cancer obtaining an area under receiver operating characteristic of 0.91. The methodology presented can be used for any type of sequential information such as any protein and nucleic acid sequence.

Introduction

Cancer is one of the leading causes of death worldwide and human colon cancer (HCC) has an important social impact.¹ HCC represents the uncontrolled growth of abnormal cells in the colon part of the intestine² due to DNA transformation/mutation. Therefore, these cells invade and destroy normal tissues around or even distant organs by spreading through the blood/lymphatic system.

The initial stage of this disease is represented by adenomatous polyps in the colon that may develop into cancer over time. The most frequent diagnosis method is the colonoscopy and the therapy consists of surgery followed by chemotherapy. If the cancer is detected early, it can be frequently cured. Even if in the last few years the rate of mortality caused by this type of cancer has decreased due to better personalized treatments and new detection methods, HCC is still very common in men and women all over the world. This disease has complex causes that include age, diet, smoking, genetic background, DNA mutations and external factors.³ The National Cancer Institute (NCI) in U.S. implemented in its web (<http://www.cancer.gov/colorectalcancerrisk/>) a colorectal cancer risk assessment tool, an interactive tool to help estimate

a person's risk of developing colorectal cancer. The tool is based on the work published in *Journal of Clinical Oncology*⁴ and it can estimate the risk for men and women who are between the ages of 50 and 85, African American, Asian American/Pacific Islander, Hispanic/Latino or White, but it cannot accurately estimate the risk for people who have problems such as ulcerative colitis, Crohn's disease, familial adenomatous polyposis (FAP), hereditary nonpolyposis colorectal cancer (HNPCC) or personal history of colorectal cancer. Therefore, the development of simple and fast theoretical methods for searching HCC biomarkers before the adenoma or in the initial stages of the disease becomes very important.

In this paper, the Quantitative Structure–Disease Relationship (QSDR)⁵ will be used, which is similar to Quantitative Structure–Activity Relationship (QSAR).^{6–13} QSDR is one of the widely used methods for predicting protein properties linked with diseases and uses macromolecular graph descriptors, named topological indices (TIs). Molecular graph theory is a branch of mathematical chemistry dedicated to encode the protein/DNA/RNA/drug information in graph representations using TIs.^{14–18} Graphical approaches for studying biological systems can provide useful insights into protein folding kinetics,¹⁹ enzyme-catalyzed reactions,^{20–23} inhibition kinetics of processive nucleic acid polymerases and nucleases,^{24–28} DNA sequence analysis,²⁹ anti-sense strands base frequencies,³⁰ analysis of codon usage,^{31,32} protein networks in parasites^{33–36} and in complicated network system research.³⁷ Graphic representation was also used to study the evolution of protein sequences³⁸ and drug metabolism systems.³⁹ Particularly, the wengxiang diagrams/graphs⁴⁰

^a Department of Information and Communications Technologies, University of A Coruña, Campus Elviña, 15071 A Coruña, Spain. E-mail: jseoane@udc.es; Fax: +34 981167160; Tel: +34 981167000 ext 1302

^b Department of Microbiology & Parasitology, Faculty of Pharmacy, University of Santiago de Compostela, 15782, Santiago de Compostela, Spain. Fax: +34 981594912; Tel: +34 981563100

were recently used to analyze the mechanism of protein–protein interactions^{41,42} and gain some very interesting insights. Interesting implementations of graph-based models for drug–protein and protein–protein interactions are presented in Bio-AIMS tools at <http://bio-aims.udc.es/TargetPred.php>.

Other interesting fields to apply the graph theory are the oncology and clinical proteomics. A classification model for discriminating prostate cancer patients from the control group with connectivity indices was constructed by González-Díaz *et al.*⁴³ Vilar's group designed a QSAR model for alignment-free prediction of HBC biomarkers based on electrostatic potentials of protein pseudofolding HP-lattice networks.⁴⁴ Prediction models for HCC using two different types of protein graphs were previously published: a HP lattice type¹⁰ and a star-graph type.⁴⁵

The current work proposes an improved cancer–non-cancer classification model for HCC based on protein square Randic spiral-graph TIs⁴⁶ obtained from protein primary sequences and Naïve Bayes classifiers.⁴⁷ Similar studies based on the spiral graph have been published: QSDR models for prostate cancer using mass spectra input data,⁵ Quantitative Proteome–Property Relationships (QPPRs) for finding biomarkers of organic drugs using blood mass spectra^{48,49} or chemical research in toxicology.⁵⁰ Naïve Bayes classifiers have been recently used for different problems such as the protein quaternary structure,⁵¹ for protein subcellular location,⁵² classification of DNA repair genes into ageing-related or non-ageing-related,⁵³ genomic data integration to reduce the misclassification rate in predicting protein–protein interactions,⁵⁴ prediction of human protein–protein interaction to explore underlying cancer-related pathway crosstalk,⁵⁵ prediction of Alzheimer's disease from genome-wide data⁵⁶ or virtual screening and chemical biology.⁵⁷

Materials and methods

The description of the methodology followed in this work is presented in Fig. 1. The input data are represented by the amino acid sequences (primary structure) of the protein related or not with HCC. By using new software programmed by our group, CULSPIN,⁵⁸ the sequences of amino acids are transformed into spiral graphs and the corresponding topological indices. The resulting numbers that characterized each graph (that is a protein graphical representation) are then used in Weka⁵⁹ to find the best QSDR classification model. The final model is used to predict if a new protein is linked with HCC using only its amino acid sequence.

Protein set

This work is based on the same datasets used in the previous studies with lattice- and star-type graphs^{10,45} for protein linked with HCC. The sets of protein primary sequences are represented by a set of 69 HCC cancer proteins⁶⁰ and 276 non-cancer proteins.^{61,62} To avoid homology bias and remove the redundant sequences from the benchmark dataset, a cut-off threshold of 25% was imposed^{63,64} to exclude those proteins from the benchmark datasets that have equal to or greater than 25% sequence identity to any other one in a same subset. However, in this study we did not use such a stringent criterion

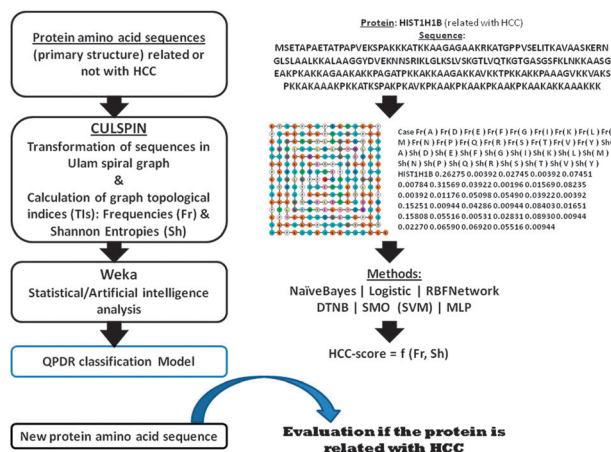


Fig. 1 Flowchart of building the QSDR classification models for HCC/non-HCC-related proteins.

because the currently available data do not allow us to do so. Otherwise, the numbers of proteins for some subsets would be too few to have statistical significance.

Ulam spiral graphs

In 1963 the mathematician Stanislaw M. Ulam discovered certain interesting aspects in relation to the disposition that adopt the prime numbers when placing the natural numbers following the shape of a spiral. Then this disposition became highly popularized as a visual picture in a number of Scientific American magazines in 1964.

To construct the spiral one must write down a regular grid of numbers, starting with one at the centre, and spiralling out the rest of integer numbers just as shown in Fig. 2A. In mathematics, this is a simple method of graphing numbers that reveals hidden patterns in numeric series and sequences. In molecular sciences this spiral representation was associated to a graph in order to represent DNA nucleotide sequences in a letter sequence of four classes (A, T, G, and C).

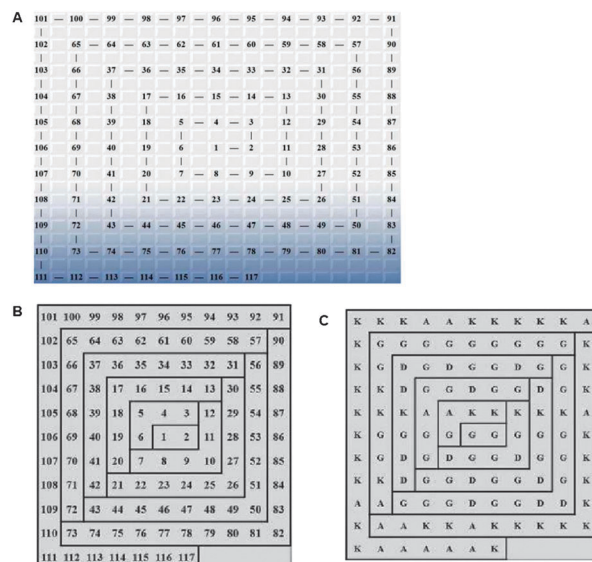


Fig. 2 Spiral of a regular grid of numbers (A), the letter gnomon division (B) and the letter gnomon division (C).

used: one considering all the available TIs and the other one including a subset of the TIs after performing feature selection.

In recent years, feature selection (FS) has become the focus of much research in areas of application for which a great amount of variables is available. Among the objectives of FS, we can consider the following as some of the most important ones: to avoid overfitting and improve model performance, to provide faster and more cost-effective models and to gain a deeper insight into the underlying processes that generated the data.⁷² In the context of classification, feature selection techniques can be organized into three categories, depending on how they combine the feature selection search with the construction of the classification model: filter methods, wrapper methods and embedded methods. In this paper, several FS techniques were applied, but the best results were obtained by combining Correlation-based Feature Subset Selection, CfsSubsetEval,⁷³ which is correlation-based and thus a filter method, with Best First, which uses hill climbing augmented with a backtracking facility or by combining Consistency-based Feature Subset Selection, ConsistencySubsetEval,⁷⁴ which is also a filter method, with Linear Forward Selection, LinearForwardSelection,⁷⁵ which is an extension of Best First. Filter methods assess the relevance of features by looking only at the intrinsic properties of the data. Feature selection has been widely used in bioinformatics.⁷⁶

Artificial Neural Networks (ANNs) have been extensively used for classification problems. In this paper, the Multilayer Perceptron (MLP) has been utilized. An MLP is a feedforward artificial neural network model that maps input data onto a set of appropriate outputs. It consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron, also known as processing element, with a nonlinear activation function. This ANN uses a supervised learning technique called back-propagation in order to train the network. As well as the MLP, Support Machine Vectors (SVM) are nonlinear classifiers. SVM induce linear separators or hyperplanes in the space of characteristics. This type of classifier has proved to be very useful when dealing with high dimensionality problems. Bayesian methods have also been applied to this type of problem. These methods are based on Bayes' theory of probability. Not only they allow performing classification, but they also allow finding relationships among attributes. Among them, we can find Naïve Bayes, which assumes that the attributes are independent. Finally, DTNB allows obtaining classification models based on "IF-THEN-ELSE" rules or on hierarchical structures such as trees.

Among the independent dataset test, sub-sampling or *k*-fold (e.g., 5 or 10-fold) cross-over test, and jackknife test, which are often used for examining the accuracy of a statistical prediction method,⁷⁷ the jackknife test was deemed the least arbitrary that can always yield a unique result for a given benchmark dataset, as elucidated and demonstrated in ref. 78. Therefore, the jackknife test has been increasingly recognized and widely adopted by investigators to test the power of various prediction methods (see, e.g., ref. 79–87). However, to reduce the computational time, 10-fold cross-validation^{88–90} has been used to verify the accuracy of the models. Hence, the original dataset is partitioned into 10 subsets. Of the 10 subsets, a single subset is retained as the validation data for testing the model and

the remaining are used as training data. The cross-validation process is then repeated 10 times, with each of the 10 subsets used exactly once as the validation data. Thus, classification accuracy percentages were calculated for the test group with the corresponding AUROCs. AUROC (Area under Receiver Operating Characteristic)⁹¹ represents the goodness of a predictor in a binary classification task and its values close to 1 show that the model has an excellent classification capacity.

Statistics

In the case of the best classification model, additional statistical studies have been presented. For this model, we calculated the sensitivity (Se), specificity (Sp), positive predictive value (PPV) and negative predictive value (NPV) for each cut-off point to evaluate the diagnostic accuracy.⁹² We also calculated the diagnostic odds ratio (DOR) which expresses the strength of the association between test result and disease: it is the ratio of the odds of a positive result in a person with the target condition compared to a person without the condition.⁹³ A DOR of 1 suggests that the test provides no diagnostic evidence. Moreover, we also calculated the likelihood ratios (LRs) which describe how many times a person with the target condition is more likely to have a particular test result than a person without that condition. LRs contribute to change the probability that a target condition is present after the test has been made. Binary tests have two LRs, positive and negative (LR+, LR-). An LR of 1 indicates no diagnostic value.

Since Naïve Bayes needs all the variables to be independent, the squared-chi test was used to ensure this condition. This analysis was performed using the PASW Statistics 18 statistical package version 18.0.0.⁹⁴

Results

More than 18 classification models were tested with the aim of finding the equation which is able to discriminate between proteins related to HCC. The initial attributes include 40 spiral graph TIs obtained with CULSPIN: 20 frequencies (Fr) and 20 Shannon entropies (Sh). Feature selection was used in order to consider the minimum number of attributes and, after that, the different classification methods were applied. Table 1 presents the classification results for the test group and the AUROC values. The classifications used only the frequencies, only the Shannon entropies and both of the TIs. These results were obtained using the Weka package.

The best QSDR classification model that can predict if a protein is HCC-related was created with the Naïve Bayes method, based only on 11 Shannon entropies of the spiral graph. The Naïve Bayes classifier estimates the probability conditioned to the class, assuming that the attributes are conditionally independent, given a class Y. This assumption can be described as follows:

$$P(\text{sh}|Y = \text{HCC}) = \prod_{i=1}^d P(\text{sh}_i|Y = \text{HCC}) \quad (3)$$

where each set of attributes $\text{Sh} = \{\text{Sh}_1, \text{Sh}_2, \text{Sh}_3, \dots, \text{Sh}_d\}$ contains *d* attributes.

Instead of computing the probability conditioned to a class for each combination of Sh, it is only necessary to estimate the conditioned probability of each Sh_i given an output Y.

Table 1 Classification scores and AUROCs for test data

Method	Fr		Sh		Both	
	Accuracy (%)	AUROC	Accuracy (%)	AUROC	Accuracy (%)	AUROC
Naïve Bayes	88.99	0.89	90.92	0.91	89.80	0.90
Logistic	82.40	0.86	83.41	0.87	86.95	0.89
RBFNetwork	88.99	0.88	89.29	0.90	88.92	0.90
DTNB	85.10	0.88	85.74	0.87	84.29	0.88
SVM	85.85	0.89	86.03	0.89	86.89	0.90
MLP	86.77	0.88	87.07	0.87	86.29	0.89

This approach does not require a large set for training in order to obtain a good estimation of the probability. To classify each test sample, the Naïve Bayes classifier calculates the posterior probability of each class Y:

$$P(\text{HCC}|\text{sh}) = \frac{P(\text{HCC}) \prod_{i=1}^4 P(\text{sh}_i|\text{HCC})}{P(\text{HCC})} \quad (4)$$

Since $P(\text{Sh})$ is the same for each output $Y = \text{HCC}$, selecting the class that maximizes the numerator is enough,

$$P(\text{HCC}) \prod_{i=1}^d P(\text{sh}_i|\text{HCC}) \quad (5)$$

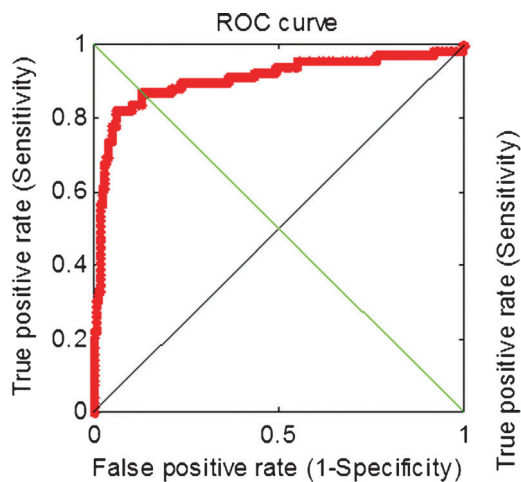
This output represents the probability of HCC, while $\text{Sh}(c)$ are the Shannon entropy topological indices of class c for the protein spiral graphs.

The model obtained a classification accuracy of 90.92% and it showed an AUROC of 0.91 (Fig. 4) for the test group. This AUROC value demonstrates that the model has excellent classification potential, by providing a very good prediction for HCC-related proteins.

The above results are typically considered as excellent in the literature QSAR/QSDR models.^{44,95–98}

Diagnostic performance

Table 2 shows diagnostic accuracy and predictive values of Naïve Bayes for two different cut-offs. These results were obtained for the HCC test group. Better values were obtained for a cut-off of 0.5. Although the specificity is lower than the one obtained for a cut-off of 0.1940, the sensitivity is higher. In addition, the NPV for a cut-off of 0.5 is 83.6, compared to 63.2 for a cut-off of 0.1940.

**Fig. 4** AUROC of Naïve Bayes for HCC.**Table 2** Diagnostic accuracy and predictive values of Naïve Bayes for HCC

Cut-off	0.1940	0.5
AUC	0.91 (0.86–0.96)	0.91 (0.86–0.96)
TP	60	51
FP	35	10
TN	241	266
FN	9	18
Se ^a	87.3 (83.4–91.2)	96.4 (94.2–98.6)
Sp ^a	87.0 (79.0–94.9)	73.9 (63.6–84.3)
PPV ^a	94.6 (94.1–98.7)	93.7 (90.8–96.5)
NPV ^a	63.2 (53.5–72.9)	83.6 (74.3–92.9)
LR ⁺ ^b	0.1	0.3
LR ⁻ ^b	6.9	20.4
DOR ^b	45.9	75.4

TP, true positive cases (correct diagnosis); FP, false positive cases (over-diagnosis); TN, true negative cases (correct diagnosis); FN, false negative cases (missed cases); Se, sensitivity; Sp, specificity; PPV, positive predictive value; NPV, negative predictive value; LR, likelihood ratio; DOR, diagnostic odds ratio. ^a Values as percentage (%) and 95% of confidence interval (95% CI). ^b Values as ratio value.

Finally, there is a great difference in terms of DOR. Therefore, it is better to consider a cut-off of 0.5.

This model obtains a great diagnostic capacity for both cut-offs. In this sense, LR⁻ is > 6 for both cut-offs, however, LR⁺ is < 1. These results confirm that the model developed here allows diagnosing the absence of HCC.

Conclusion

This study proposes a new classification model for HCC using the spiral graph TIs of the protein amino acid sequences. The best model based on only 11 Shannon entropy TIs and obtained with the Naïve Bayes method proves the excellent predictive ability (90.92%) for new proteins linked with HCC. Previous works have proposed different models for HCC based on topological indices of star and lattice graphs for the same dataset.

The star graph-based study⁴⁵ proposed an input-coded multi-target classification model for two types of cancer, human breast cancer (HBC) and human colon cancer (HCC). The general discriminant analysis method generated the best model with the training/predicting set accuracies of 90.0% for the forward stepwise model type. The model was based on 5 pure and mixed star graph TIs obtained with S2SNet software.⁹⁹

The other study using the same protein dataset is based on lattice graphs.¹⁰ 69 proteins related to HCC and a control group of 200 proteins non-related to HCC were represented through an HP Lattice type Network. Starting from the generated graphs a set of descriptors of electrostatic potential

type has been calculated. The Linear Discriminant Analysis (LDA) helped to establish a QSAR model of relatively high percentage of good classification (between 80% and 90%) to differentiate between HCC and non-HCC proteins.

Therefore, the current study proposes an alternative model with better prediction capacity, based on a different type of protein graph, on Shannon entropy information of the graph and on a simple statistical method such as Naïve Bayes.

This work can help in oncology proteomics or serve as model for other studies, for proteins linked with different diseases. In addition, the new CULSPIN application is demonstrating its capacity to transform simple protein sequences into TIs and to be the base of protein studies.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors,¹⁰⁰ we shall make efforts in our future work to provide a web-server for the method presented in this paper.

Acknowledgements

Cristian R. Munteanu and González-Díaz H. acknowledge the funding support for a research position by the “Isidro Parga Pondal” program from Xunta de Galicia and the European Social Fund (ESF). The work of Vanessa Aguiar-Pulido is supported by the “Plan I2C” program, from Xunta de Galicia, and by the ESF. This work is supported by the following projects: RD07/0067/0005 funded by the Carlos III Health and 10SIN105004PR funded by Economy and Industry Department of Xunta de Galicia.

References

- 1 A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray and M. J. Thun, *Ca-Cancer J. Clin.*, 2008, **58**, 71–96.
- 2 B. Boursi and N. Arber, *Ca-Cancer J. Clin.*, 2007, **13**, 2274–2282.
- 3 C. Schafmayer, S. Buch, J. H. Egberts, A. Franke, M. Brosch, A. El Sharawy, M. Conring, M. Koschnick, S. Schwiedernoch, A. Katalinic, B. Kremer, U. R. Folsch, M. Krawczak, F. Fandrich, S. Schreiber, J. Tepel and J. Hampe, *Int. J. Cancer*, 2007, **121**, 555–558.
- 4 A. N. Freedman, M. L. Slattey, R. Ballard-Barbash, G. Willis, B. J. Cann, D. Pee, M. H. Gail and R. M. Pfeiffer, *J. Clin. Oncol.*, 2009, **27**, 686–693.
- 5 G. Ferino, H. Gonzalez-Diaz, G. Delogu, G. Podda and E. Uriarte, *Biochem. Biophys. Res. Commun.*, 2008, **372**, 320–325.
- 6 A. Tropsha, *Mol. Inf.*, 2010, **29**, 476–488.
- 7 K. Roy and I. Mitra, *Comb. Chem. High Throughput Screening*, 2011, **14**, 450–474.
- 8 E. Demchuk, P. Ruiz, S. Chou and B. A. Fowler, *Toxicol. Appl. Pharmacol.*, 2011, **254**, 192–197.
- 9 J. Devillers and A. T. Balaban, *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach, The Netherlands, 1999.
- 10 S. Vilar, H. Gonzalez-Diaz, L. Santana and E. Uriarte, *J. Theor. Biol.*, 2009, **261**, 449–458.
- 11 H. Wei, C. H. Wang, Q. S. Du, J. Meng and K. C. Chou, *Med. Chem.*, 2009, **5**, 305–317.
- 12 J. Wang, X. Y. Wang, M. Shu, Y. Q. Wang, Y. Lin, L. Wang, X. M. Cheng and Z. H. Lin, *Protein Pept. Lett.*, 2011, **18**, 956–963.
- 13 X. Hou, J. Du, H. Fang and M. Li, *Protein Pept. Lett.*, 2011, **18**, 440–449.
- 14 O. Ivanciuc, T. Ivanciuc, D. Cabrol-Bass and A. T. Balaban, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 631–643.
- 15 M. Randić and A. T. Balaban, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 532–539.
- 16 M. Randić, J. Zupan and M. Novic, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1339–1344.
- 17 M. Randić, J. Zupan and D. Vikić-Topić, *J. Mol. Graphics Modell.*, 2007, **26**, 290–305.
- 18 C. R. Munteanu, E. Fernandez-Blanco, J. A. Seoane, P. Izquierdo-Novo, J. A. Rodriguez-Fernandez, J. M. Prieto-Gonzalez, J. R. Rabunal and A. Pazos, *Curr. Pharm. Des.*, 2010, **16**, 2640–2655.
- 19 K. C. Chou, *Biophys. Chem.*, 1990, **35**, 1–24.
- 20 K. C. Chou, *J. Biol. Chem.*, 1989, **264**, 12074–12079.
- 21 K. C. Chou and S. Forsen, *Biochem. J.*, 1980, **187**, 829–835.
- 22 K. C. Chou and W. M. Liu, *J. Theor. Biol.*, 1981, **91**, 637–654.
- 23 P. Kuzmic, K. Y. Ng and T. D. Heath, *Anal. Biochem.*, 1992, **200**, 68–73.
- 24 I. W. Althaus, J. J. Chou, A. J. Gonzales, M. R. Diebel, K. C. Chou, F. J. Kezdy, D. L. Romero, P. A. Aristoff, W. G. Tarpley and F. Reusser, *Biochemistry*, 1993, **32**, 6548–6554.
- 25 I. W. Althaus, J. J. Chou, A. J. Gonzales, M. R. Diebel, K. C. Chou, F. J. Kezdy, D. L. Romero, P. A. Aristoff, W. G. Tarpley and F. Reusser, *J. Biol. Chem.*, 1993, **268**, 6119–6124.
- 26 I. W. Althaus, J. J. Chou, A. J. Gonzales, R. J. LeMay, M. R. Deibel, K. C. Chou, F. J. Kezdy, D. L. Romero, R. C. Thomas and P. A. Aristoff and, *et al, Experientia*, 1994, **50**, 23–28.
- 27 I. W. Althaus, K. C. Chou, R. J. Lemay, K. M. Franks, M. R. Deibel, F. J. Kezdy, L. Resnick, M. E. Busso, A. G. So, K. M. Downey, D. L. Romero, R. C. Thomas, P. A. Aristoff, W. G. Tarpley and F. Reusser, *Biochem. Pharmacol.*, 1996, **51**, 743–750.
- 28 K. C. Chou, F. J. Kezdy and F. Reusser, *Anal. Biochem.*, 1994, **221**, 217–230.
- 29 X. Q. Qi, J. Wen and Z. H. Qi, *J. Theor. Biol.*, 2007, **249**, 681–690.
- 30 K. C. Chou, C. T. Zhang and D. W. Elrod, *J. Protein Chem.*, 1996, **15**, 59–61.
- 31 K. C. Chou and C. T. Zhang, *AIDS Res. Hum. Retroviruses*, 1992, **8**, 1967–1976.
- 32 C. T. Zhang and K. C. Chou, *J. Mol. Biol.*, 1994, **238**, 1–8.
- 33 Y. Rodriguez-Soca, C. R. Munteanu, J. Dorado, J. Rabuñal, A. Pazos and H. González-Díaz, *Polymer*, 2010, **51**, 264–273.
- 34 H. Gonzalez-Diaz, L. Muino, A. M. Anadon, F. Romaris, F. J. Prado-Prado, C. R. Munteanu, J. Dorado, A. P. Sierra, M. Mezo, M. Gonzalez-Warleta, T. Garate and F. M. Ubeira, *Mol. BioSyst.*, 2011, **7**, 1938–1955.
- 35 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. Garíjo, J. Sansano and F. M. Ubeira, *J. Proteome Res.*, 2011, **10**, 1698–1718.
- 36 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 and F. M. Ubeira, *J. Theor. Biol.*, 2011, **276**, 229–249.
- 37 P. Riera-Fernández, C. R. Munteanu, N. Pedreira-Souto, R. Martín-Romalde, A. Duardo-Sanchez and H. González-Díaz, *Curr. Bioinf.*, 2011, **6**, 94–121.
- 38 Z. C. Wu, X. Xiao and K. C. Chou, *J. Theor. Biol.*, 2010, **267**, 29–34.
- 39 K. C. Chou, *Curr. Drug Metab.*, 2010, **11**, 369–378.
- 40 K. C. Chou, W. Z. Lin and X. Xiao, *Nat. Sci.*, 2011, **3**, 862–865 (openly accessible at <http://www.scirp.org/journal/NS/>).
- 41 G. P. Zhou, *J. Theor. Biol.*, 2011, **284**, 142–148.
- 42 G. P. Zhou, *Protein Pept. Lett.*, 2011, **18**, 966–978.
- 43 H. González-Díaz, G. Ferino, G. Podda and E. Uriarte, *Electron. Conf. Synth. Org. Chem.*, 2007, **11**(G1:1), 10.
- 44 S. Vilar, H. Gonzalez-Diaz, L. Santana and E. Uriarte, *J. Comput. Chem.*, 2008, **29**, 2613–2622.
- 45 C. R. Munteanu, A. L. Magalhaes, E. Uriarte and H. Gonzalez-Diaz, *J. Theor. Biol.*, 2009, **257**, 303–311.
- 46 M. Randić, N. Lers, D. Plavšić, S. Basak and A. T. Balaban, *Chem. Phys. Lett.*, 2005, **407**, 205–208.
- 47 A. Y. Ng and M. I. Jordan, *Adv. Neural Inf. Process. Syst.*, 2002, **2**, 841–848.
- 48 M. Cruz-Monteagudo, C. R. Munteanu, F. Borges, M. N. Cordeiro, E. Uriarte and H. Gonzalez-Diaz, *Bioorg. Med. Chem.*, 2008, **16**, 9684–9693.
- 49 M. Cruz-Monteagudo, C. R. Munteanu, F. Borges, M. N. Cordeiro, E. Uriarte, K. C. Chou and H. González-Díaz, *Polymer*, 2008, **49**, 5575–5587.

- 50 M. Cruz-Monteagudo, H. Gonzalez-Diaz, F. Borges, E. R. Dominguez and M. N. Cordeiro, *Chem. Res. Toxicol.*, 2008, **21**, 619–632.
- 51 P. Mitra and D. Pal, *Structure*, 2011, **19**, 304–312.
- 52 C. Jackson, E. Glory-Afshar, R. F. Murphy and J. Kovacevic, *Bioinformatics (Oxford, England)*, 2011, **27**, 1854–1859.
- 53 A. A. Freitas, O. Vasieva and J. P. de Magalhaes, *BMC Genomics*, 2011, **12**, 27.
- 54 C. Xing and D. B. Dunson, *PLoS Comput. Biol.*, 2011, **7**, e1002110.
- 55 Y. Xu, W. Hu, Z. Chang, H. Duanmu, S. Zhang, Z. Li, Z. Li, L. Yu and X. Li, *J. R. Soc., Interface*, 2011, **8**, 555–567.
- 56 W. Wei, S. Visweswaran and G. F. Cooper, *J. Am. Med. Inf. Assoc.*, 2011, **18**, 370–375.
- 57 A. Bender, *Methods Mol. Biol. (Totowa, N. J.)*, 2011, **672**, 175–196.
- 58 L. G. Pérez Montoto, F. J. Prado-Prado, C. R. Munteanu and H. González Díaz, CULSPIN . Compute ULam SPiral INdices, Santiago de Compostela, 2009.
- 59 M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. A. Witten, *SIGKDD Explor.*, 2009, **11**, 10–18.
- 60 T. Sjoblom, S. Jones, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, D. Mandelker, R. J. Leary, J. Ptak, N. Silliman, S. Szabo, P. Buckhaults, C. Farrell, P. Meeh, S. D. Markowitz, J. Willis, D. Dawson, J. K. Willson, A. F. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B. H. Park, K. E. Bachman, N. Papadopoulos, B. Vogelstein, K. W. Kinzler and V. E. Velculescu, *Science*, 2006, **314**, 268–274.
- 61 P. D. Dobson, Y. D. Cai, B. J. Stapley and A. J. Doig, *Curr. Med. Chem.*, 2004, **11**, 2135–2142.
- 62 P. D. Dobson and A. J. Doig, *J. Mol. Biol.*, 2005, **345**, 187–199.
- 63 K. C. Chou and H. B. Shen, *Anal. Biochem.*, 2007, **370**, 1–16.
- 64 K. C. Chou and H. B. Shen, *PLoS One*, 2010, **5**, e9931.
- 65 N. Rappin and R. Dunn, *wxPython in Action*, Manning Publications Co, Greenwich, CT, 2006.
- 66 P. Langley, W. Iba and K. Thompson, *An analysis of Bayesian classifiers*, San Jose, CA, 1992.
- 67 T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, 2001.
- 68 J. Moody and C. J. Darken, *Neural Comput.*, 1989, **1**, 281–294.
- 69 M. Hall and E. Frank, presented in part at the In Proceedings of 21st Florida Artificial Intelligence Research Society Conference, Miami, Florida, 2008.
- 70 V. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, New York, USA, 1998.
- 71 C. Bishop, *Neural Networks for pattern recognition*, Oxford University Press, New York, 1995.
- 72 I. Guyon and A. Elisseeff, *J. Mach. Learn. Res.*, 2003, **3**, 1157–1182.
- 73 M. A. Hall and L. A. Smith, *Correlation-based Feature Subset Selection for Machine Learning*, Hamilton, New Zealand, 1998.
- 74 H. Liu and R. Setiono, presented in part at the 13th International Conference on Machine Learning, 1996.
- 75 M. Guetlein, E. Frank, M. Hall and A. Karwath, presented in part at the In Proceedings of IEEE Symposium on Computational Intelligence and Data Mining, 2009.
- 76 Y. Saeyns, I. Inza and P. Larranaga, *Bioinformatics (Oxford, England)*, 2007, **23**, 2507–2517.
- 77 K. C. Chou and C. T. Zhang, *Crit. Rev. Biochem. Mol. Biol.*, 1995, **30**, 275–349.
- 78 K. C. Chou, *J. Theor. Biol.*, 2011, **273**, 236–247.
- 79 C. Chen, L. Chen, X. Zou and P. Cai, *Protein Pept. Lett.*, 2009, **16**, 27–31.
- 80 M. Esmaeili, H. Mohabatkar and S. Mohsenzadeh, *J. Theor. Biol.*, 2010, **263**, 203–209.
- 81 D. N. Georgiou, T. E. Karakasidis, J. J. Nieto and A. Torres, *J. Theor. Biol.*, 2009, **257**, 17–26.
- 82 Z. C. Wu, X. Xiao and K. C. Chou, *Mol. BioSyst.*, 2011, **7**, 3287–3297.
- 83 H. Mohabatkar, M. Mohammad Beigi and A. Esmaeili, *J. Theor. Biol.*, 2011, **281**, 18–23.
- 84 H. Mohabatkar, *Protein Pept. Lett.*, 2010, **17**, 1207–1214.
- 85 K. C. Chou, Z. C. Wu and X. Xiao, *PLoS One*, 2011, **6**, e18258.
- 86 K. C. Chou, Z. C. Wu and X. Xiao, *Mol. BioSyst.*, 2012, **8**, 629–641.
- 87 X. Xiao, P. Wang and K. C. Chou, *Mol. BioSyst.*, 2011, **7**, 911–919.
- 88 G. J. McLachlan, K.-A. Do and C. Ambrose, *Analyzing Microarray Gene Expression Data*, Wiley-Interscience, Hoboken, New Jersey, 2004.
- 89 R. Kohavi, *A study of cross-validation and bootstrap for accuracy estimation and model selection*, Montreal, Quebec, Canada, 1995.
- 90 R. Picard and D. Cook, *J. Am. Stat. Assoc.*, 1984, **79**, 575–583.
- 91 J. A. Hanley and B. J. McNeil, *Radiology*, 1982, **143**, 29–36.
- 92 K. Linnet, *Clin. Chem.*, 1988, **34**, 1379–1386.
- 93 A. S. Glas, J. G. Lijmer, M. H. Prins, G. J. Bonsel and P. M. Bossuyt, *J. Clin. Epidemiol.*, 2003, **56**, 1129–1135.
- 94 SPSS, SPSS, Chicago, 2009.
- 95 Y. Marrero-Ponce, H. G. Diaz, V. R. Zaldivar, F. Torrens and E. A. Castro, *Bioorg. Med. Chem.*, 2004, **12**, 5331–5342.
- 96 A. H. Morales, M. A. Cabrera Perez and M. P. Gonzalez, *J. Mol. Model*, 2006, **12**, 769–780.
- 97 E. Estrada and E. Molina, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 791–797.
- 98 J. A. Castillo-Garit, Y. Marrero-Ponce, F. Torrens, R. Garcia-Domenech and V. Romero-Zaldivar, *J. Comput. Chem.*, 2008, **29**, 2500–2512.
- 99 C. R. Munteanu and H. González-Díaz, S2SNet – Sequence to Star Network, Santiago de Compostela, Spain, 2008.
- 100 K. C. Chou and H. B. Shen, *Nat. Sci.*, 2009, **1**, 63–92.

Trypano-PPI: A Web Server for Prediction of Unique Targets in Trypanosome Proteome by using Electrostatic Parameters of Protein–protein Interactions

Yamilet Rodríguez-Soca,[†] Cristian R. Munteanu,[‡] Julián Dorado,[‡] Alejandro Pazos,[‡]
 Francisco J. Prado-Prado,[†] and Humberto González-Díaz^{*,†}

Department of Microbiology & Parasitology, Faculty of Pharmacy, University of Santiago de Compostela, 15782, Santiago de Compostela, Spain, and Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruña, Campus de Elviña, 15071, A Coruña, Spain

Received September 15, 2009

Abstract: *Trypanosoma brucei* causes African trypanosomiasis in humans (HAT or African sleeping sickness) and Nagana in cattle. The disease threatens over 60 million people and uncounted numbers of cattle in 36 countries of sub-Saharan Africa and has a devastating impact on human health and the economy. On the other hand, *Trypanosoma cruzi* is responsible in South America for Chagas disease, which can cause acute illness and death, especially in young children. In this context, the discovery of novel drug targets in Trypanosome proteome is a major focus for the scientific community. Recently, many researchers have spent important efforts on the study of protein–protein interactions (PPIs) in pathogen Trypanosome species concluding that the low sequence identities between some parasite proteins and their human host render these PPIs as highly promising drug targets. To the best of our knowledge, there are no general models to predict Unique PPIs in Trypanosome (TPPIs). On the other hand, the 3D structure of an increasing number of Trypanosome proteins is reported in databases. In this regard, the introduction of a new model to predict TPPIs from the 3D structure of proteins involved in PPI is very important. For this purpose, we introduced new protein–protein complex invariants based on the Markov average electrostatic potential $\xi_k(R_i)$ for amino acids located in different regions (R_i) of i -th protein and placed at a distance k one from each other. We calculated more than 30 different types of parameters for 7866 pairs of proteins (1023 TPPIs and 6823 non-TPPIs) from more than 20 organisms, including parasites and human or cattle hosts. We found a very simple linear model that predicts above 90% of TPPIs and non-TPPIs both in training and independent test subsets using only two parameters. The parameters were $^d\xi_k(s) = |\xi_k(s_1) - \xi_k(s_2)|$, the absolute difference between the $\xi_k(s_i)$ values

on the surface of the two proteins of the pairs. We also tested nonlinear ANN models for comparison purposes but the linear model gives the best results. We implemented this predictor in the web server named TrypanoPPI freely available to public at <http://miaja.tic.udc.es/Bio-AIMS/TrypanoPPI.php>. This is the first model that predicts how unique a protein–protein complex in Trypanosome proteome is with respect to other parasites and hosts, opening new opportunities for antitrypanosome drug target discovery.

Keywords: *Trypanosoma* proteome • African trypanosomiasis • Chagas disease • Markov chains • protein–protein interactions • 3D-electrostatic potential • protein surface • machine learning • artificial neural networks

Introduction

African trypanosomiasis is a vector-borne parasitic disease caused by protozoan parasites of the *Trypanosoma* genus. *Trypanosoma brucei* species can infect both humans and animals, causing Human African Trypanosomiasis (HAT, also known as African sleeping sickness) in man and Nagana in cattle. The disease threatens over 60 million people and uncounted numbers of cattle in 36 countries of sub-Saharan Africa and has a devastating impact on human health and the economy in affected areas. Unless treated, HAT is always fatal. Political instability and economic problems are leading factors for the reduced efficacy in vector and disease control, resulting in a resurgence of disease that continues to this day (<http://www.who.int/tdr>). On the other hand, *Trypanosoma cruzi* is responsible in South America for Chagas disease, which can cause acute illness and death, especially in young children. More commonly, patients develop a chronic form of the disease that affects most organs of the body, often causing fatal damage to the heart and digestive tract. Transmission occurs via bloodsucking triatomine bugs and congenitally from mother to the unborn child but can also occur through contaminated blood transfusions (<http://www.who.int/en/>).¹ Control of HAT relies primarily on chemotherapy. Nevertheless, there is a very limited arsenal of drugs, but they generally have shortcomings, such as high toxicity and emerging resistance. The drugs currently available to treat HAT have been available for more than half a century. Early stages of HAT are treated with

* To whom correspondence should be addressed. H. González-Díaz: Faculty of Pharmacy, USC, Spain. Phone: +34-981-563100. Fax: +34-981 594912. E-mail: humberto.gonzalez@usc.es or gonzalezdiaz@yahoo.es.

[†] University of Santiago de Compostela.

[‡] University of A Coruña.

Trypano-PPI

pentamidine and suramin. Side effects for both drugs are significant and the failure rate is high, especially for suramin. Late stages of HAT can be treated with melarsoprol, a melaminophenyl arsenical compound that is able to cross the blood-brain barrier. Drug-induced side effects are severe and up to 5% of those patients treated die. The only alternative to melarsoprol is eflornithine, an analogue of ornithine that acts as an inhibitor of trypanosomal ornithine decarboxylase, leading to a block in polyamine synthesis. Side effects are significant but eflornithine is much less toxic than melarsoprol. However, eflornithine is not effective against the form of the disease caused by *T. brucei rhodesiense* in East Africa. In this context, a research aimed at the identification and validation of novel drug targets is a major goal for the scientific community.¹

Recently, many researchers have spent important efforts on the experimental and/or theoretical studies of protein–protein interactions (PPIs) in pathogen Trypanosoma species.^{2–4} In addition, the knowledge about the biology of these parasites according to the investigation of PPIs may guide researchers on the search of new drug targets for HAT or Chagas disease. For instance, Choe and Moyersoen et al. carried out the analysis of the sequence motifs responsible for the interactions of peroxins 14 and 5, which are involved in glycosome biogenesis in *Trypanosoma brucei*. Glycosome biogenesis in trypanosomatids occurs via a process that is homologous to peroxisome biogenesis in other eukaryotes. Glycosomal matrix proteins are synthesized in the cytosol and imported post-translationally. The import process involves a series of PPIs starting from recognition of glycosomal matrix proteins by a receptor in the cytosol. Most proteins to be imported contain so-called PTS-1 or PTS-2 targeting sequences recognized by the receptor proteins PEX5 and PEX7, respectively. These authors measured the strength of the interactions between *Trypanosoma brucei* PEX14 and PEX5 by a fluorescence assay, using (i) a panel of N-terminal regions of TbPEX14 protein variants and (ii) a series of different peptides derived from TbPEX5, each containing one of the three WXXXF/Y motifs present in this receptor protein. They concluded that the low sequence identities of PEX14 and PEX5 between parasite and its human host, and the vital importance of proper glycosome biogenesis to the parasite, render these peroxins highly promising drug targets.⁵

These types of results indicate that unique PPIs of Trypanosoma parasites (TPPIs) and not presented in humans may be promising targets for the development of safe drugs with low toxicity. In addition, the high number of possible PPIs in parasite and human hosts makes difficult, in terms of time and resources, the exhaustive experimental investigation. It determines that, not only in parasites but in all organisms in general, the development of predictive models for PPIs becomes a very useful tool to guide the discovery of new drug targets.⁶ In general, there are many structural parameters and theoretical methods that are useful in proteome research for protein–structure function relationship studies. In principle, the same type of methods may be used for the prediction of PPIs in humans and other organisms. Many of them use sequence alignment techniques, phylogenic techniques, or alignment-free parameters to construct and/or analyze proteins or PPIs in terms of protein networks representations (as input or output of the analysis).^{7–15} Sequence only methods are often faster than 3D ones and need less structural information. On the contrary, 3D methods give a more clear idea on the structure of the protein and may be used to predict proteins with known

spatial structure but unknown function.^{16–24} Alignment-free methods involve topological indices, signal analysis, or 3D structural parameters; see for instance the works of Giuliani, Zbilut, Kirshnan, Torrens, Marrero-Ponce, Caballero and Fernandez, Estrada, Ivanciuc and others.^{25–36} The importance of these last methods is that these functionally nonannotated structures are becoming common in the Protein Data Bank (PDB) with the development of powerful characterization techniques.³⁷ Specifically, in this work, we are interested in computational methods predicting TPPIs that determine the formation of a noncovalent complex between the two proteins that can be isolated and the 3D structure chemically characterized as a potential drug target. Protein complexes are essential in order to be able to understand principles of cellular organizations. As the sizes of PPI networks are increasing, accurate and fast protein complex prediction from these PPI networks can serve as a guide for biological experiments to discover novel protein complexes.³⁸ Otherwise, it is the direct prediction of complexes by protein–protein docking but it may become computationally expensive if we aim at performing the screening of large databases.³⁹ In addition, with the introduction of Internet, the development of new predictive methods has become the first step in the application of computational techniques to proteome research. Nowadays, it is not enough to develop a fast and accurate predictive model, we should also implement it into public servers, preferably of free access, for the use of the scientific community. The server packages developed by Chou and Shen to predict the function of proteins from structural parameters or explore protein structures^{40–43} are good examples in this regard. These may be used by proteome research scientists by interacting with user-friendly interfaces. It means that the user does not need to be an expert on the theoretical details behind this kind of model including the vast literature published by Chou et al. on the development of models with pseudo amino acid composition parameters or the use of machine learning classification techniques and other algorithms.^{44–48} In any case, to the best of our knowledge, in the literature there is no theoretical method to predict unique TPPIs in Trypanosome proteome that are not present in humans or other organisms, based on the 3D structure of the two proteins involved in the interaction.

Separately, González-Díaz et al. introduced the method called Markovian chemicals in silico design (MARCH-INSIDE 1.0) for the computational design of small-sized drugs. The approach uses a Markov chain model (MCM) of the intramolecular movement of electrons to calculate structural parameters of drugs. In subsequent studies, we have extended this method to perform a fast calculation of 2D and 3D alignment-free structural parameters based on molecular vibrations in RNA secondary structures, or electrostatic potential, and van der Waals interactions in proteins. Currently, the method was renamed as Markov chains invariants for networks simulation and design (MARCH-INSIDE 2.0). This describe more adequately the broad uses of the method that describes the structure of drugs,⁴⁹ RNA,⁵⁰ and proteins,^{51–53} as well as drug–drug networks⁵⁴ and drug–protein interactions.⁵⁵ The MARCH-INSIDE may be used also to study PPIs, bacteria–bacteria coaggregation, parasite–host interactions, and other systems with a MCM associated to a network. In very recent reviews, we have discussed the last applications of this method.^{7,56,57} For all these reasons, in

this work we use MARCH-INSIDE approach to solve the problem of predicting specific TPPIs from the 3D structure of the two proteins involved. Last, we implement the first public server for prediction of TPPIs.

Methods

Electrostatic Parameters of Protein–Protein Interaction.

In previous works, we used 3D-electrostatic potential invariants derived with an MCM to describe the 3D structure of one protein backbone in structure–property relationship studies. The parameters used $\xi_k(R)$ to represent the average electrostatic potential (ξ) due to the interactions between all pairs of amino acids (*aa*). The chosen amino acids are those with the electrostatic charges q_i and q_j that are allocated inside a specific protein region (R) and placed one from each other at a distance d_{ij} equal to or shorter than k -times the cutoff distance (see details in previous works).^{53,58–61} In this work, we want to use $\xi_k(R)$ values of two proteins, $\xi_k^1(R)$ for protein 1 and $\xi_k^2(R)$ for protein 2, to generate structural parameters describing PPI between these proteins. To this end, we introduce here for the first time a new type of PPI invariants in the sense that they do not depend on the interchange between proteins in such a way that we do not need to label and distinguish them for calculation. We introduce, with this objective, three types of invariants: PPI electrostatic average invariant $\xi_k(R)$, PPI electrostatic absolute-difference invariant, and PPI electrostatic product invariant:

$${}^a\xi_k^1(R_1, {}^2R_1) = \frac{1}{2}[\xi_k^1(R_1) + \xi_k^2(R_1)] \quad (1)$$

$${}^d\xi_k^1(R_1, {}^2R_1) = |\xi_k^1(R_1) - \xi_k^2(R_1)| \quad (2)$$

$${}^p\xi_k^1(R_1, {}^2R_1) = \xi_k^1(R_1) \cdot \xi_k^2(R_1) \quad (3)$$

Notably, to guarantee that these parameters are invariants to protein labeling as 1 or 2, we have to use always the same ${}^1R = {}^2R = R$ and $k_1 = k_2 = k$ values. To calculate the $\xi_k(R)$ values for each protein the method uses as a source of protein macromolecular descriptors the stochastic matrices ${}^1\Pi_e$ built up as a squared matrices ($n \times n$), where n is the number of *aa* in the protein. The subscript e points to the electrostatic type of molecular force field. The method considers a hypothetical situation in which every j^{th} -*aa* has general potential ξ_j isolated in the space. All these potentials can be listed as elements of the vector ${}^0\varphi_r$. It can be supposed that, after this initial situation, all the amino acids interact with the energy ${}^1E_{ij}$ with every other aa_j in the protein. For the sake of simplicity, a truncation function α_{ij} is applied in such a way that a short-term interaction takes place in a first approximation only between neighboring amino acids ($\alpha_{ij} = 1$ if $d_{ij} \leq$ cutoff distance). Otherwise, the interaction is banished ($\alpha_{ij} = 0$). Neglecting direct interactions between distant *aa* in ${}^1\Pi_e$ does not avoid the possibility that potential interactions propagate between those *aa* within the protein backbone in an indirect manner. Consequently, in the present model long-range electrostatic interactions are allowed (not forbidden) but estimated indirectly using the natural powers of ${}^n\Pi_e = ({}^1\Pi_e)^n$. The use of MCM theory allows a simple and fast model to calculate the average values of ξ_k considering indirect interaction between any aa_i and the other aa_j after previous interaction of aa_j with other k neighbor amino acids. As follows, we give the general formula for any potential and specific formulas as well:⁷

$$\xi_k(R) = \sum_{j=1 \in R}^n {}^A p_k(j) \cdot \left(\frac{q_j}{d_{j0}} \right) = \sum_{j=1 \in R}^n {}^A p_k(j) \cdot \varphi_j = {}^0\pi_e^T \cdot {}^k\Pi_e \cdot {}^0\varphi_e = {}^0\pi_e^T \cdot ({}^1\Pi_e)^k \cdot {}^0\varphi_e \quad (4)$$

It is remarkable that the average general potentials ξ_k depend on the absolute probabilities ${}^A p_k(j)$ with which the amino acids interact with other amino acids and their k -order. The potential $\xi_k(R)$ depends also on the initial unperturbed potential of the amino acid $\varphi_j = (q_j/d_{j0})$; with d_{j0} equal to the distance from the carbon C_α of the amino acid to the center of the protein ($x, y, z) = (0, 0, 0)$. In the equations presented above, the ${}^A p_k(j)$ values are calculated with the vector of absolute initial probabilities, ${}^0\pi_r$, and the matrix ${}^1\Pi_e$ based on the Chapman-Kolmogorov equations. In particular, the evaluation of such expansions for $k = 0$ gives the initial average unperturbed electrostatic potential (ξ_0), for $k = 1$ the short-range potential (ξ_1), for $k = 2$ the middle-range potential (ξ_2), and for $k = 3$ the long-range one. This expansion is illustrated for the tripeptide Ala-Val-Trp (AVW):⁷

$$\xi_0 = [{}^A p_0(A), {}^A p_0(V), {}^A p_0(W)] \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \varphi_A \\ \varphi_V \\ \varphi_W \end{bmatrix} = {}^A p_0(A) \cdot \varphi_A + {}^A p_0(V) \cdot \varphi_V + {}^A p_0(W) \cdot \varphi_W \quad (5)$$

$$\xi_1 = [{}^A p_0(A), {}^A p_0(V), {}^A p_0(W)] \cdot \begin{bmatrix} {}^1 p_{AA} & {}^1 p_{AV} & 0 \\ {}^1 p_{VA} & {}^1 p_{VV} & {}^1 p_{VW} \\ 0 & {}^1 p_{WV} & {}^1 p_{WW} \end{bmatrix} \cdot \begin{bmatrix} \varphi_A \\ \varphi_V \\ \varphi_W \end{bmatrix} = {}^A p_1(A) \cdot \varphi_A + {}^A p_1(V) \cdot \varphi_V + {}^A p_1(W) \cdot \varphi_W \quad (6)$$

$$\xi_2 = [{}^A p_0(A), {}^A p_0(V), {}^A p_0(W)] \cdot \begin{bmatrix} {}^1 p_{AA} & {}^1 p_{AV} & 0 \\ {}^1 p_{VA} & {}^1 p_{VV} & {}^1 p_{VW} \\ 0 & {}^1 p_{WV} & {}^1 p_{WW} \end{bmatrix} \cdot \begin{bmatrix} \varphi_A \\ \varphi_V \\ \varphi_W \end{bmatrix} \quad (7)$$

$$\xi_3 = [{}^A p_0(A), {}^A p_0(V), {}^A p_0(W)] \cdot \begin{bmatrix} {}^1 p_{AA} & {}^1 p_{AV} & 0 \\ {}^1 p_{VA} & {}^1 p_{VV} & {}^1 p_{VW} \\ 0 & {}^1 p_{WV} & {}^1 p_{WW} \end{bmatrix} \cdot \begin{bmatrix} \varphi_A \\ \varphi_V \\ \varphi_W \end{bmatrix} \quad (8)$$

In order to carry out the calculations referred to in eqs 1 for any kind of potential and detailed in the previous equations, for electrostatic potential the elements (${}^1 p_{ij}$) of ${}^1\Pi_e$ and the absolute initial probabilities ${}^A p_k(j)$ were calculated as follows:⁷

$${}^1 p_{ij} = \frac{\alpha_{ij} \cdot E_{ij}}{\sum_{m=1}^{\delta+1} \alpha_{im} \cdot E_{im}} = \frac{\alpha_{ij} \cdot \frac{q_i \cdot q_j}{d_{ij}^2}}{\sum_{m=1}^{\delta+1} \alpha_{im} \cdot \frac{q_i \cdot q_m}{d_{im}^2}} \quad (9)$$

$${}^A p_0(j) = \frac{q_j}{d_{j0}} \cdot \sum_{m=1}^n \frac{q_m}{d_{0m}} \quad (10)$$

where, q_i and q_j are the AMBER electronic charge parameters⁶² for amino acids i^{th} -*aa* and the j^{th} -*aa* and the neighborhood

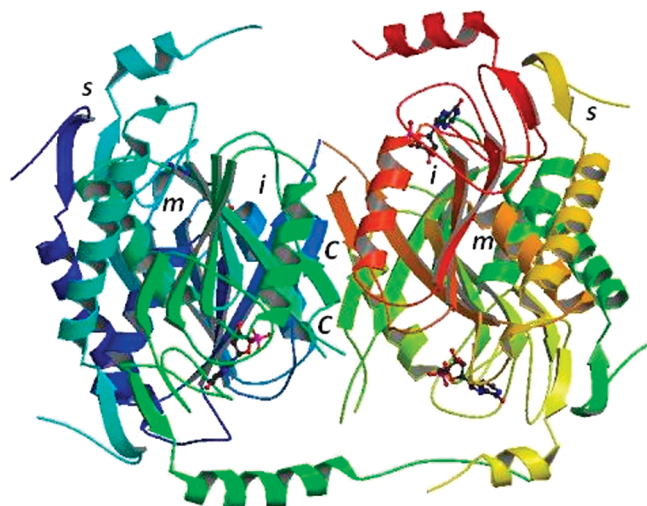


Figure 1. Example of spatial distribution of core, inner, middle, and surface amino acids.

relationship (truncation function $\alpha_{ij} = 1$) was turned on if these amino acids participate in a peptidic hydrogen bond or $d_{ij} < d_{\text{cutoff}} = 1/2(v_{\text{dw}}r_i + v_{\text{dw}}r_j)$, which is the semisum of the van der Waals radii for both *aa*.⁷ In this regard, truncation of the molecular field is usually applied to simplify all the calculations in large biological systems. The distance d_{ij} is the Euclidean distance between the C_{α} atoms of the two amino acids and d_{0j} the distance between the amino acid and the center of charge of the protein. Both kinds of distances were derived from the x , y and z coordinates of the amino acids collected from the protein PDB files. All calculations were carried out with our in-house software MARCH-INSIDE. For calculation, all water molecules and metal ions were removed.⁶³ For the calculation, the MARCH-INSIDE software divided the protein into four orbits (R) called *c*, *i*, *m* and *s* that constitute specific groups or collections of amino acids placed at the protein core (*c*), inner (*i*), middle (*m*) or surface region (*s*) (see Figure 1). The diameters of the orbits, as a percentage of the longer distance with respect to the center of charge, are 0–25 for orbit *c*, 25.1–50 for orbit *i*, 50.1–75 for orbit *m*, and 75.1–100 for orbit *s*. Figure 2 presents the flowchart of the present method.

Artificial Neural Network (ANN) Analysis. Artificial neural networks (ANN) have been used to test a linear model not based on assumptions of parametric distribution of data and nonlinear models as well. The ANNs have been trained with the software STATISTICA 6.0, for which our laboratory holds rights of use. The classification problem was solved with the *Intelligent Problem Solver* analysis by using a selection of a subset of the independent variables. The retained networks were selected by using the balance performance against diversity. Several types of ANNs have been tested such as test of the linear ANN (LNN), probabilistic neural network (PNN),⁶⁴ general regression neural network (GRNN),⁶⁵ radial basis functions (RBF),⁶⁶ and the three and four layer perceptron (Multi-Layer Perceptron, MLP).⁶⁷ The number of tested hidden units had the values of 1–1967 for RBF and 1–10 for the layer 2 of the three layer MLP and layers 2 and 3 of the four layer MLP. The linear models (LNN) are MLP without hidden neurons. The bias neurons have not been considered. The minimum classification loss threshold was 1 and the classification output encoding was entropy-based. The training algorithms were back-propagation^{68–70} (in phase one with 100 epochs and learning rate of 0.01) and conjugate gradient

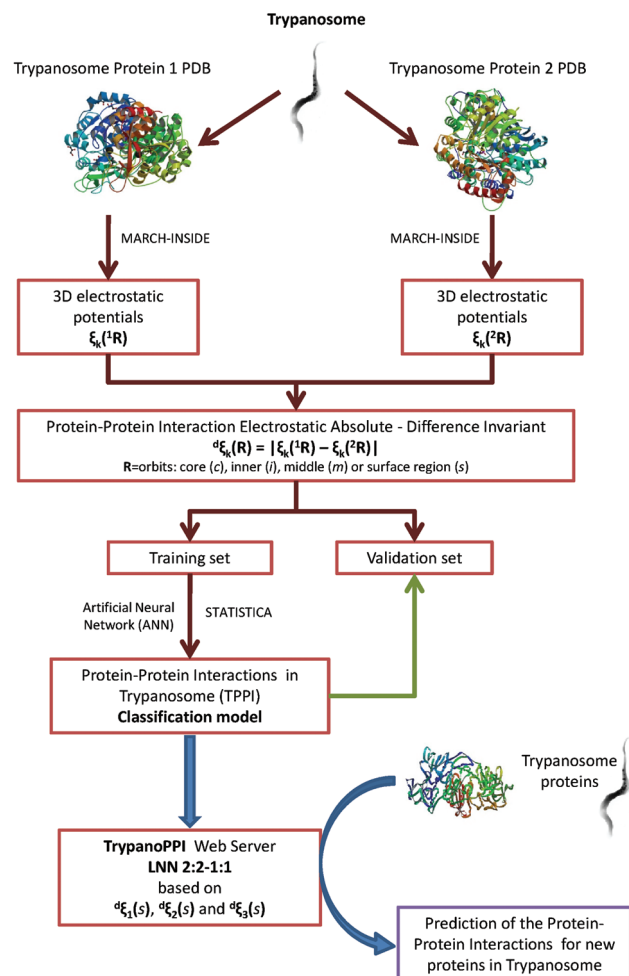


Figure 2. General scheme of work with all steps necessary to develop or use the present model.

descent⁷¹ (in phase two with 500 epochs). All the ANNs have been tested for one step (one training period); see for instance the work of Vilar et al. with ANNs.⁷² In Figure 3 we illustrate the graph representation of some of the ANNs trained in this work.

Data Set. The protein structures were downloaded from PDB⁷³ using the following schemes for PDB-database search: (i) introducing as input parameter the name of the parasite specie (*Trypanosome*) in the search item called source organism (for positive cases) or (ii) introducing the PDB IDs for all the proteins contained in the list reported in the article of Dobson and Doig.⁷⁴ The positive cases (TPPI) are those protein–protein pairs that form stable complex that have been structurally characterized (3D structure) in *Trypanosome* species. The list of negative cases (non-TPPI), search scheme (b), contains enzymes and other protein complexes present in humans and many other organisms including other parasites (see Figure 4) that are not present in *Trypanosome* species. The data set was composed by 7866 pairs of proteins (1023 TPPIs and 6823 non-TPPIs) from more than 20 organisms, including parasites and human or cattle hosts. Detailed information about the PDB ID, the values of the electrostatic potential indices, the corresponding observed classification, and the predicted classification for each TPPI or non-TPPI pair are given in the Supporting Information.

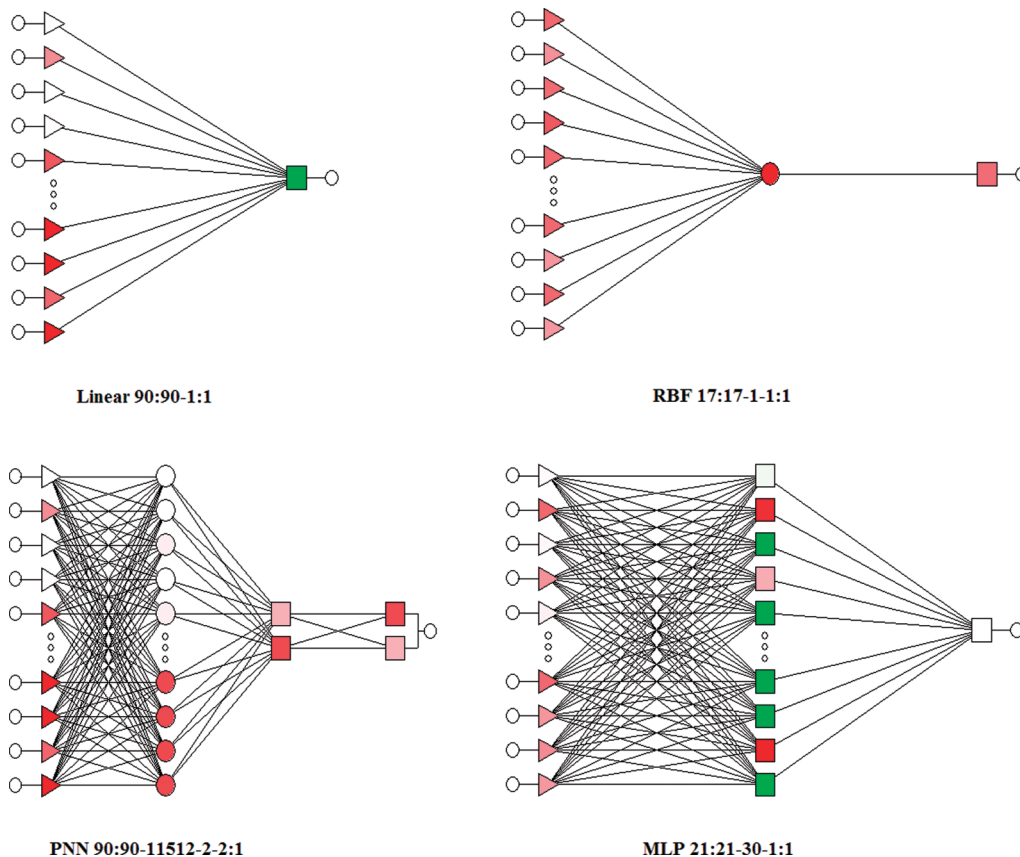


Figure 3. Illustrative examples of the topology used for some of the ANN models trained.

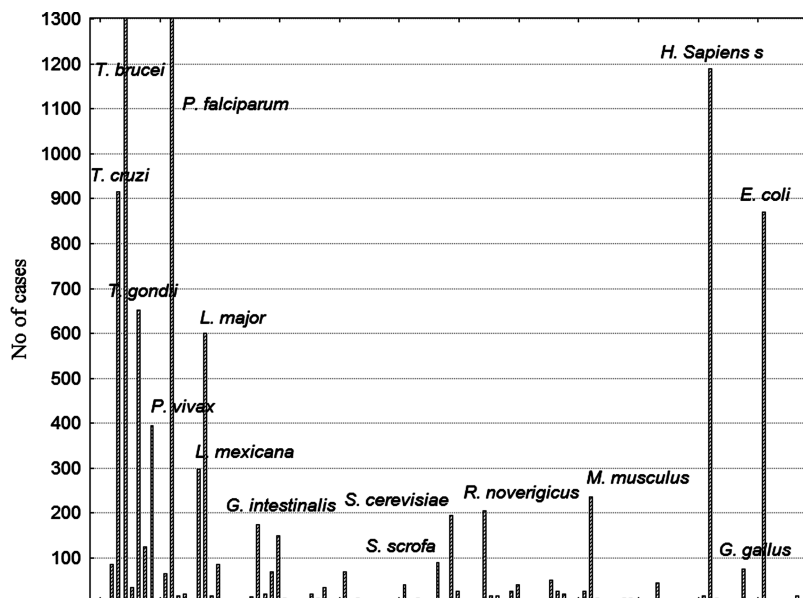


Figure 4. Histogram of number of PPI and non-PPI cases studied by organism (the order of organism in the x-axis is by first time of appearance in the list of Supporting Information).

Results and Discussion

Several researchers have demonstrated the high performance of different types of computational classifiers in structure–function relationship studies ranging from low-weight molecules to protein–protein complexes and based on different algorithms; see for instance the works of Ivanciuc about Machine Learning,^{75–78} or the works of Cai and Chou et al.^{44–48,79–81} with different classifiers. In

particular, the Linear Neural Network (LNN) algorithm, the simpler type of ANN, was used here to train different linear models based on different combinations of parameters. Table 1 depicts the results for the best models found. The profile of the ANN model was specified with a simple notation as follows: ANN type $N_{iv}:N_{in}-N_{H1}-N_{H2}-N_{on}:N_{ov}$. The ANN types presented, in addition to LNN, are multi-layer perceptron (MLP), probabilistic neural network (PNN), and radial basis

Table 1. Summary of ANN Analysis Results for Some Models

ANN profile	set	parameter	value (%)	group	TPPI	non-TPPI
LNN 2:2-1:1	Train	Sensitivity	88.2	TPPI	677	91
		Specificity	89.7	non-TPPI	526	4578
		Accuracy	89.5			
	Test	Sensitivity	91.4	TPPI	233	22
		Specificity	90.9	non-TPPI	159	1580
		Accuracy	90.9			
LNN 3:3-1:1	Train	Sensitivity	88.3	TPPI	678	90
		Specificity	89.1	non-TPPI	554	4550
		Accuracy	89.0			
	Test	Sensitivity	91.8	TPPI	234	21
		Specificity	90.5	non-TPPI	165	1574
		Accuracy	90.7			
PNN 3:3-5872-2-2:1	Train	Sensitivity	0.0	TPPI	0	768
		Specificity	100.0	non-TPPI	0	5104
		Accuracy	86.9			
	Test	Sensitivity	0.0	TPPI	0	255
		Specificity	100.0	non-TPPI	0	1739
		Accuracy	87.2			
MLP 1:1-6-5-1:1	Train	Sensitivity	88.4	TPPI	679	89
		Specificity	88.9	non-TPPI	564	4540
		Accuracy	88.9			
	Test	Sensitivity	91.8	TPPI	234	21
		Specificity	90.3	non-TPPI	168	1571
		Accuracy	90.5			
MLP 1:1-4-1:1	Train	Sensitivity	88.3	TPPI	678	90
		Specificity	88.9	non-TPPI	567	4537
		Accuracy	88.8			
	Test	Sensitivity	91.8	TPPI	234	21
		Specificity	90.5	non-TPPI	166	1573
		Accuracy	90.6			
MLP 1:1-6-1:1	Train	Sensitivity	88.4	TPPI	679	89
		Specificity	88.9	non-TPPI	564	4540
		Accuracy	88.9			
	Test	Sensitivity	91.8	TPPI	234	21
		Specificity	90.3	non-TPPI	168	1571
		Accuracy	90.5			
RBF 1:1-1-1:1	Train	Sensitivity	11.7	TPPI	90	678
		Specificity	11.3	non-TPPI	4528	576
		Accuracy	11.3			
	Test	Sensitivity	8.2	TPPI	21	234
		Specificity	9.8	non-TPPI	1569	170
		Accuracy	9.6			

function (RBF).⁸² The parameter N_{iv} is the number of input variables, N_{in} is the number of input neurons (one per input variable), N_{H1} is the number of neurons in the first Hidden layer (H1), N_{H2} is the number of neurons in the second Hidden layer (H1), N_{on} is the number of output neurons, and N_{ov} is the number of output variables. The automatically selection of variables (features) was activated for all models. Interestingly, three variables, $d\xi_1(s)$, $d\xi_2(s)$ and $d\xi_3(s)$, out of more than 30 parameters calculated appear in many models and are chosen by an additional LDA variable selection. These parameters have the general formula $d\xi_k(s) = |\xi_k(s)_{prot1} - \xi_k(s)_{prot2}|$; which are the absolute difference between the electrostatic potential values $\xi_k(s)$ for amino acids on the surface of the two proteins forming the PPI pairs. This fact indicates that the difference between the surface electrostatic potential is very important not only for PPI interactions in general but also to discriminate unique complex present in *Trypanosome* (TPPIs) and not in other organisms.

In particular, the model LNN 2:2-1:1 is the simplest model found with higher levels of sensitivity = 88.2, specificity = 89.7, and accuracy = 89.5 in training set. These values are

excellent considering that this predictor uses only two molecular descriptors of the PPI pair. The fitting of this large data set of 768 TPPIs and 5104 non-TPPIs is a very complex process from a chemical point of view. The profile 2:2-1:1 indicates that this model assigns the values of only two input variables to two input neurons that perform a weighted sum and assign the result to one output neuron, which gives the final result of classification of the case according to the threshold value that have been optimized. In addition, the model LNN 2:2-1:1 presented also higher levels of sensitivity = 91.4, specificity = 90.9, and accuracy = 90.9 in the external test set (see Table 1). We also validated the model by means of a ROC curve⁸³ analysis (see Figure 5). The values of the area under the ROC curve for this model are 0.95 and 0.96 very close to 1 (the highest possible value) and notably different from 0.5 (the value typical of a random classifier).

The comparison of linear and nonlinear models is essential to test how directly our parameters are correlated to the biological property.⁸⁴ This first search points to a linear instead of nonlinear relationship between TPPI prediction and $d\xi_k(s)$ values, giving additional proof of the validity of

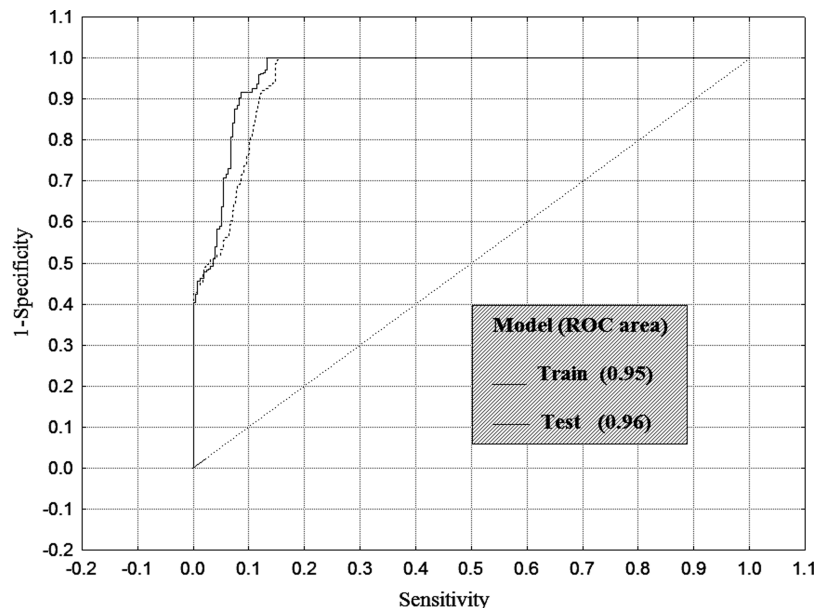


Figure 5. ROC curve for the TPPI predictor with profile LNN 1:2-1:1.

our methodology. For instance, in Table 1 we can see that more complicated models with very nonlinear profiles do not improve the linear model and sometimes give even worse results. All the models are using as input only the three variables $d\xi_1(s)$, $d\xi_2(s)$, and $d\xi_3(s)$ selected before with an LDA variable selection model. The big number of hidden neuron in the PNNs is automatically generated by the default algorithm from STATISTICA. Last, we should consider that with the advent of the Internet it is important not only to develop new predictive models for proteome research but also to carry out the implementation of these models in public web servers available to other research groups.^{40–43,59,85–88} In this regard, we have implemented this predictor at a web server freely available to public at <http://miaja.tic.udc.es/Bio-AIMS/TrypanoPPI.php>. This is the first model and web server that predicts how unique a protein–protein complex in Trypanosome proteome is with respect to other parasites and host breaking new ground for antitrypanosome drug target discovery.

Conclusions

In this paper we introduce a new type of parameters to numerically characterize protein structure in PPI studies. We also demonstrate that it is possible to distinguish between protein–protein complexes unique in *Trypanosome* species (TPPIs cases) and not present in other organisms with a linear classifier based on the absolute difference between 3D protein surface electrostatic potentials of the pair proteins. The model was implemented in a public web server, available to the scientific community for free of charge use.

Acknowledgment. We sincerely thank the kind attention and valuable comments received from both the editor Prof. Martin W. McIntosh and the unknown referee. H.G.-D. and C.R.M. acknowledge research contract sponsored by Xunta de Galicia (grant: Isidro Parga Pondal Program). We also thank partial financial support from the General Directorate of Scientific and Technologic Promotion of the Galician University System, Xunta de Galicia (grants:

2007/127 and 2007/144), and Carlos III Health Institute (grants: PIO52048 and RD07/0067/0005).

Supporting Information Available: Detailed information about the PDB ID, the values of the electrostatic potential indices, the corresponding observed classification, and the predicted classification for each TPPI or non-TPPI pair. This material is available free of charge via the Internet at <http://pubs.acs.org>.

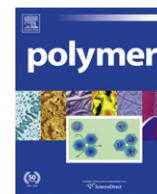
References

- (1) Naula, C.; Parsons, M.; Mottram, J. C. Protein kinases as drug targets in trypanosomes and Leishmania. *Biochim. Biophys. Acta* **2005**, *1754* (1–2), 151–9.
- (2) Cribb, P.; Serra, E. One- and two-hybrid analysis of the interactions between components of the *Trypanosoma cruzi* spliced leader RNA gene promoter binding complex. *Int. J. Parasitol.* **2009**, *39* (5), 525–32.
- (3) Juri Ayub, M.; Smulski, C. R.; Nyambega, B.; Bercovich, N.; Masiga, D.; Vazquez, M. P.; Aguilar, C. F.; Levin, M. J. Protein-protein interaction map of the *Trypanosoma cruzi* ribosomal P protein complex. *Gene* **2005**, *357* (2), 129–36.
- (4) Caro, F.; Bercovich, N.; Atorrasagasti, C.; Levin, M. J.; Vazquez, M. P. Protein interactions within the TcZFP zinc finger family members of *Trypanosoma cruzi*: implications for their functions. *Biochem. Biophys. Res. Commun.* **2005**, *333* (3), 1017–25.
- (5) Choe, J.; Moyersoan, J.; Roach, C.; Carter, T. L.; Fan, E.; Michels, P. A.; Hol, W. G. Analysis of the sequence motifs responsible for the interactions of peroxins 14 and 5, which are involved in glycosome biogenesis in *Trypanosoma brucei*. *Biochemistry* **2003**, *42* (37), 10915–22.
- (6) Chou, K. C.; Cai, Y. D. Predicting protein–protein interactions from sequences in a hybridization space. *J. Proteome Res.* **2006**, *5* (2), 316–22.
- (7) González-Díaz, H.; González-Díaz, Y.; Santana, L.; Ubeira, F. M.; Uriarte, E. Proteomics, networks and connectivity indices. *Proteomics* **2008**, *8*, 750–778.
- (8) Wu, J.; Mellor, J. C.; DeLisi, C. Deciphering protein network organization using phylogenetic profile groups. *Genome Inform. Ser. Workshop Genome Inform.* **2005**, *16* (1), 142–9.
- (9) McDermott, J.; Samudrala, R. Enhanced functional information from predicted protein networks. *Trends Biotechnol.* **2004**, *22* (2), 60–2, discussion 62–3.
- (10) Huynen, M. A.; Snel, B.; von Mering, C.; Bork, P. Function prediction and protein networks. *Curr. Opin. Cell Biol.* **2003**, *15* (2), 191–8.
- (11) Jeong, H.; Mason, S. P.; Barabasi, A. L.; Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **2001**, *411* (6833), 41–2.

- (12) Carmi, S.; Levanon, E. Y.; Havlin, S.; Eisenberg, E. Connectivity and expression in protein networks: proteins in a complex are uniformly expressed. *Phys. Rev. E: Stat. Nonlin. Soft Matter Phys.* **2006**, *73* (3 Pt 1), 031909.
- (13) Bornholdt, S.; Schuster, H. G. *Handbook of Graphs and Complex Networks: From the Genome to the Internet*; WILEY-VCH GmbH & CO. KGa.: Weinheim, 2003.
- (14) Estrada, E. Protein bipartivity and essentiality in the yeast protein-protein interaction network. *J. Proteome Res.* **2006**, *5* (9), 2177–84.
- (15) Estrada, E. Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics* **2006**, *6* (1), 35–40.
- (16) Sharon, I.; Davis, J. V.; Yona, G. Prediction of protein-protein interactions: a study of the co-evolution model. *Methods Mol. Biol.* **2009**, *541*, 61–88.
- (17) Liu, L.; Cai, Y.; Lu, W.; Feng, K.; Peng, C.; Niu, B. Prediction of protein-protein interactions based on PseAA composition and hybrid feature selection. *Biochem. Biophys. Res. Commun.* **2009**, *380* (2), 318–22.
- (18) Skrabanek, L.; Saini, H. K.; Bader, G. D.; Enright, A. J. Computational prediction of protein-protein interactions. *Mol. Biotechnol.* **2008**, *38* (1), 1–17.
- (19) Najafabadi, H. S.; Salavati, R. Sequence-based prediction of protein-protein interactions by means of codon usage. *Genome Biol.* **2008**, *9* (5), R87.
- (20) Kim, S.; Shin, S. Y.; Lee, I. H.; Kim, S. J.; Sriram, R.; Zhang, B. T. PIE: an online prediction system for protein-protein interactions from text. *Nucleic Acids Res.* **2008**, *36* (Web Server issue), W411–5.
- (21) Jaeger, S.; Gaudan, S.; Leser, U.; Rebholz-Schuhmann, D. Integrating protein-protein interactions and text mining for protein function prediction. *BMC Bioinf.* **2008**, *9* (Suppl 8), S2.
- (22) Burger, L.; van Nimwegen, E. Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Mol. Syst. Biol.* **2008**, *4*, 165.
- (23) Scott, M. S.; Barton, G. J. Probabilistic prediction and ranking of human protein-protein interactions. *BMC Bioinf.* **2007**, *8*, 239.
- (24) Ivanciuc, O.; Schein, C. H.; Braun, W. Data mining of sequences and 3D structures of allergenic proteins. *Bioinformatics* **2002**, *18* (10), 1358–64.
- (25) Fernández, M.; Caballero, J.; Fernández, L.; Abreu, J. I.; Garriga, M. Protein radial distribution function (P-RDF) and Bayesian-Regularized Genetic Neural Networks for modeling protein conformational stability: Chymotrypsin inhibitor 2 mutants. *J. Mol. Graph Model* **2007**, *26* (4), 748–759.
- (26) Fernández, L.; Caballero, J.; Abreu, J. I.; Fernández, M. Amino Acid Sequence Autocorrelation Vectors and Bayesian-Regularized Genetic Neural Networks for Modeling Protein Conformational Stability: Gene V Protein Mutants. *Proteins* **2007**, *67*, 834–852.
- (27) Fernández, M.; Caballero, F.; Fernández, L.; Abreu, J. I.; Acosta, G. Classification of conformational stability of protein mutants from 3D pseudo-folding graph representation of protein sequences using support vector machines. *Proteins* **2008**, *70* (1), 167–175.
- (28) Zbilut, J. P.; Giuliani, A.; Colosimo, A.; Mitchell, J. C.; Colafranceschi, M.; Marwan, N.; Webber, C. L., Jr.; Uversky, V. N. Charge and hydrophobicity patterning along the sequence predicts the folding mechanism and aggregation of proteins: a computational approach. *J. Proteome Res.* **2004**, *3* (6), 1243–53.
- (29) Krishnan, A.; Giuliani, A.; Zbilut, J. P.; Tomita, M. Network scaling invariants help to elucidate basic topological principles of proteins. *J. Proteome Res.* **2007**, *6* (10), 3924–34.
- (30) Krishnan, A.; Zbilut, J. P.; Tomita, M.; Giuliani, A. Proteins as networks: usefulness of graph theory in protein science. *Curr. Protein Pept. Sci.* **2008**, *9* (1), 28–38.
- (31) Giuliani, A.; Benigni, R.; Zbilut, J. P.; Webber, C. L., Jr.; Sirabella, P.; Colosimo, A. Nonlinear signal analysis methods in the elucidation of protein sequence-structure relationships. *Chem. Rev.* **2002**, *102* (5), 1471–92.
- (32) Marrero-Ponce, Y.; Medina-Marrero, R.; Castillo-Garrit, J. A.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. Protein linear indices of the 'macromolecular pseudograph alpha-carbon atom adjacency matrix' in bioinformatics. Part I: prediction of protein stability effects of a complete set of alanine substitutions in Arc repressor. *Bioorg. Med. Chem.* **2005**, *13* (8), 3003–15.
- (33) Marrero-Ponce, Y.; Medina-Marrero, R.; Castro, E. A.; Ramos de Armas, R.; González-Díaz, H.; Romero-Zaldivar, V.; Torrens, F. Protein Quadratic Indices of the "Macromolecular Pseudograph's α -Carbon Atom Adjacency Matrix". 1. Prediction of Arc Repressor Alanine-mutant's Stability. *Molecules* **2004**, *9*, 1124–1147.
- (34) Estrada, E.; Uriarte, E.; Vilar, S. Effect of Protein Backbone Folding on the Stability of Protein-Ligand Complexes. *J. Proteome Res.* **2006**, *5*, 105–111.
- (35) Ivanciuc, O.; Braun, W. Robust quantitative modeling of peptide binding affinities for MHC molecules using physical-chemical descriptors. *Protein Pept. Lett.* **2007**, *14* (9), 903–16.
- (36) Ivanciuc, O.; Oezguen, N.; Mathura, V. S.; Schein, C. H.; Xu, Y.; Braun, W. Using property based sequence motifs and 3D modeling to determine structure and functional regions of proteins. *Curr. Med. Chem.* **2004**, *11* (5), 583–93.
- (37) von Grothuss, M.; Plewczynski, D.; Ginalski, K.; Rychlewski, L.; Shakhnovich, E. I. PDB-UF: database of predicted enzymatic functions for unannotated protein structures from structural genomics. *BMC Bioinf.* **2006**, *7*, 53.
- (38) Chua, H. N.; Ning, K.; Sung, W. K.; Leong, H. W.; Wong, L. Using indirect protein-protein interactions for protein complex prediction. *J. Bioinform. Comput. Biol.* **2008**, *6* (3), 435–66.
- (39) Smith, G. R.; Sternberg, M. J. Prediction of protein-protein interactions by docking methods. *Curr. Opin. Struct. Biol.* **2002**, *12* (1), 28–35.
- (40) Shen, H. B.; Chou, K. C. PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.* **2008**, *373* (2), 386–8.
- (41) Shen, H. B.; Chou, K. C. Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. *Protein Eng. Des. Sel.* **2007**, *20* (11), 561–7.
- (42) Chou, K. C.; Shen, H. B. MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun.* **2007**, *360* (2), 339–45.
- (43) Chou, K. C.; Shen, H. B. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* **2008**, *3* (2), 153–62.
- (44) Chou, K. C. Prediction of G-protein-coupled receptor classes. *J. Proteome Res.* **2005**, *4* (4), 1413–8.
- (45) Chou, K. C.; Elrod, D. W. Bioinformatical analysis of G-protein-coupled receptors. *J. Proteome Res.* **2002**, *1* (5), 429–33.
- (46) Chou, K. C.; Elrod, D. W. Prediction of enzyme family classes. *J. Proteome Res.* **2003**, *2* (2), 183–90.
- (47) Chou, K. C.; Shen, H. B. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *J. Proteome Res.* **2006**, *5*, 1888–97.
- (48) Chou, K. C.; Shen, H. B. Large-scale predictions of Gram-negative bacterial protein subcellular locations. *J. Proteome Res.* **2006**, *5*, 3420–8.
- (49) 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.
- (50) González-Díaz, H.; de Armas, R. R.; Molina, R. Markovian negentropies in bioinformatics. 1. A picture of footprints after the interaction of the HIV-1 Psi-RNA packaging region with drugs. *Bioinformatics* **2003**, *19* (16), 2079–87.
- (51) Agüero-Chapin, G.; Varona-Santos, J.; de la Riva, G. A.; Antunes, A.; Gonzalez-Villa, T.; Uriarte, E.; Gonzalez-Diaz, H. Alignment-Free Prediction of Polygalacturonases with Pseudofolding Topological Indices: Experimental Isolation from *Coffea arabica* and Prediction of a New Sequence. *J. Proteome Res.* **2009**, *8* (4), 2122–28.
- (52) González-Díaz, 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* (2), 904–08.
- (53) Concu, R.; Dea-Ayuela, M. A.; Perez-Montoto, L. G.; Bolas-Fernandez, F.; Prado-Prado, F. J.; Podda, G.; Uriarte, E.; Ubeira, F. M.; Gonzalez-Diaz, H. Prediction of Enzyme Classes from 3D Structure: A General Model and Examples of Experimental-Theoretic Scoring of Peptide Mass Fingerprints of *Leishmania* Proteins. *J. Proteome Res.* **2009**, *8* (9), 4372–82.
- (54) Santana, L.; Gonzalez-Diaz, H.; Quezada, E.; Uriarte, E.; Yanez, M.; Vina, D.; Orallo, F. Quantitative structure-activity relationship and complex network approach to monoamine oxidase A and B inhibitors. *J. Med. Chem.* **2008**, *51* (21), 6740–51.
- (55) Vina, D.; Uriarte, E.; Orallo, F.; Gonzalez-Diaz, H. Alignment-Free Prediction of a Drug-Target Complex Network Based on Parameters of Drug Connectivity and Protein Sequence of Receptors. *Mol. Pharm.* **2009**, *6* (3), 825–35.
- (56) 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* (18), 1676–90.
- (57) González-Díaz, H.; Vilar, S.; Santana, L.; Uriarte, E. Medicinal Chemistry and Bioinformatics - Current Trends in Drugs Discovery with Networks Topological Indices. *Curr. Top. Med. Chem.* **2007**, *7* (10), 1025–39.

- (58) Concu, R.; Podda, G.; Uriarte, E.; Gonzalez-Diaz, H. 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.* **2009**, *30*, 1510–20.
- (59) 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. *Biochim. Biophys. Acta* **2009**, *1794* (12), 1784–94.
- (60) González-Díaz, H.; Saíz-Urra, L.; Molina, R.; Uriarte, E. Stochastic molecular descriptors for polymers. 2. Spherical truncation of electrostatic interactions on entropy based polymers 3D-QSAR. *Polymer* **2005**, *46*, 2791–8.
- (61) Gonzalez-Diaz, H.; Molina, R.; Uriarte, E. Recognition of stable protein mutants with 3D stochastic average electrostatic potentials. *FEBS Lett.* **2005**, *579* (20), 4297–301.
- (62) Liu, Y.; Beveridge, D. L. Exploratory studies of ab initio protein structure prediction: multiple copy simulated annealing, AMBER energy functions, and a generalized born/solvent accessibility solvation model. *Proteins* **2002**, *46* (1), 128–46.
- (63) González-Díaz, H.; Sanchez-Gonzalez, A.; Gonzalez-Diaz, Y. 3D-QSAR study for DNA cleavage proteins with a potential anti-tumor ATCUN-like motif. *J Inorg Biochem* **2006**, *100* (7), 1290–7.
- (64) Speckt, D. F. Probabilistic Neural Networks. *Neural Networks* **1990**, *3* (1), 109–18.
- (65) Caudill, M. GRNN and Bear It. *AI Expert* **1993**, *8* (5), 28–33.
- (66) Buhmann, M. D. *Radial Basis Functions: Theory and Implementations*; Cambridge University Press: New York, 2003.
- (67) Haykin, S. *Neural Networks: A Comprehensive Foundation*, 2nd ed.; Prentice Hall: New York, 1998.
- (68) Patterson, D. *Artificial Neural Networks*; Prentice Hall: Singapore, 1996.
- (69) Bryson, A. E.; Ho, Y.-C. *Applied optimal control: optimization, estimation, and control*; Blaisdell Publishing Company or Xerox College Publishing: Waltham, MA, 1969.
- (70) Haykin, S. *Neural Networks: A Comprehensive Foundation*; Macmillan Publishing: New York, 1994.
- (71) Bishop, C. *Neural Networks for Pattern Recognition*; University Press: Oxford, 1995.
- (72) 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.
- (73) Ivanisenko, V. A.; Pintus, S. S.; Grigorovich, D. A.; Kolchanov, N. A. PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Res.* **2005**, *33* (Database issue), D183–7.
- (74) Dobson, P. D.; Doig, A. J. Distinguishing enzyme structures from non-enzymes without alignments. *J. Mol. Biol.* **2003**, *330* (4), 771–83.
- (75) Ivanciuc, O. Weka machine learning for predicting the phospholipidosis inducing potential. *Curr. Top. Med. Chem.* **2008**, *8* (18), 1691–709.
- (76) Ivanciuc, O. Drug Design with Machine Learning. In *Encyclopedia of Complexity and Systems Science*, Meyers, R. A., Ed.; Springer-Verlag: Berlin, 2009; pp 2159–96.
- (77) Ivanciuc, O. , Drug Design with Artificial Neural Networks. In *Encyclopedia of Complexity and Systems Science*, Meyers, R. A. , Ed. Springer-Verlag: Berlin, 2009; pp 2139–59.
- (78) Ivanciuc, O. , Drug Design with Artificial Intelligence Methods. In *Encyclopedia of Complexity and Systems Science*, Meyers, R. A., Ed.; Springer-Verlag: Berlin, 2009; pp 2113–39.
- (79) Cai, Y. D.; Chou, K. C. Using functional domain composition to predict enzyme family classes. *J. Proteome Res.* **2005**, *4* (1), 109–11.
- (80) Cai, Y. D.; Chou, K. C. Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. *J. Proteome Res.* **2005**, *4* (3), 967–71.
- (81) Chou, K. C.; Shen, H. B. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.* **2007**, *6*, 1728–1734.
- (82) Rabow, A. A.; Scheraga, H. A. Lattice neural network minimization. Application of neural network optimization for locating the global-minimum conformations of proteins. *J. Mol. Biol.* **1993**, *232* (4), 1157–68.
- (83) Hill, T.; Lewicki, P. *STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining*; StatSoft: Tulsa, 2006; Vol. 1, p 813.
- (84) 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* (12), 4137–50.
- (85) Schlessinger, A.; Yachdav, G.; Rost, B. PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics* **2006**, *22* (7), 891–3.
- (86) Mewes, H. W.; Frishman, D.; Mayer, K. F.; Munsterkotter, M.; Noubibou, O.; Pagel, P.; Rattei, T.; Oesterheld, M.; Ruepp, A.; Stumpflen, V. MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.* **2006**, *34* (Database issue), D169–72.
- (87) Xie, D.; Li, A.; Wang, M.; Fan, Z.; Feng, H. LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Res.* **2005**, *33* (Web Server issue), W105–10.
- (88) McDermott, J.; Guerquin, M.; Frazier, Z.; Chang, A. N.; Samudrala, R. BIOVERSE: enhancements to the framework for structural, functional and contextual modeling of proteins and proteomes. *Nucleic Acids Res.* **2005**, *33* (Web Server issue), W324–5.

PR900827B



Plasmod-PPI: A web-server predicting complex biopolymer targets in plasmodium with entropy measures of protein–protein interactions

Yamilet Rodriguez-Soca^a, Cristian R. Munteanu^b, Julian Dorado^b, Juan Rabuñal^b, Alejandro Pazos^b, Humberto González-Díaz^{a,*}

^a Department of Microbiology & Parasitology, Faculty of Pharmacy, 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

ARTICLE INFO

Article history:

Received 18 October 2009

Received in revised form

7 November 2009

Accepted 12 November 2009

Available online 26 November 2009

Keywords:

Protein–Protein interactions

Plasmodium proteome

Protein 3D-Electrostatic interactions

ABSTRACT

We can define structural indices of polymer or biopolymer complex structures and use them in the prediction of new drug targets in parasites. For instance, *Plasmodium falciparum* causes the most severe form of Malaria and kills up to 2.7 million people annually whereas *Plasmodium vivax* is geographically the most widely distributed cause with more than 80 million clinical cases. Due to drug resistance and toxicity, discovering novel drug targets is mandatory; such as Protein–Protein Complexes unique in this pathogen and not present in human host (pPPCs). Additionally, the 3D structure of an increasing number of Plasmodium proteins is being reported in public databases making easier the development of bio-informatics models to predict pPPCs. In addition, some PPCs expressed both in parasite and human, such as DHFR synthase, play a significant role in drug resistance in both Malaria and Human Cancer. However, there are no general models to predict pPPCs using indices of PPC biopolymer structure. Therefore, we introduced herein new Markov Chain numerical descriptors of protein–protein Interactions (PPIs) based on electrostatic entropy measures and calculated these parameters for 5257 pairs of proteins (774 pPPCs and 4483 non-pPPCs) from more than 20 organisms, including parasite and human hosts. We found a simple Classification Tree with high Accuracy, Sensitivity, and Specificity (90.2–98.5%) both in training and independent test sub-sets and implemented this predictor in the user-friendly web server PlasmodPPI freely available at <http://miaja.tic.udc.es/Bio-AIMS/PlasmodPPI.php>.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

Plasmodium falciparum (*P. falciparum*) represents one of the strongest selective forces on the human genome. This stable and perennial pressure has contributed to the progressive accumulation in the exposed populations of genetic adaptations to malaria. Descriptive genetic epidemiology provides the initial step of a logical procedure of consequential phases spanning from the identification of genes involved in the resistance/susceptibility to diseases, to the determination of the underlying mechanisms and finally to the possible translation of the acquired knowledge in new control tools [1]. In addition, *Plasmodium vivax* (*P. vivax*) is geographically the most widely distributed cause of malaria in people, with up to 2.5 billion people at risk and an estimated 80 million to 300 million clinical cases every year, including severe

disease and death. Despite this large burden of disease, *P. vivax* is overlooked and left in the shadow of the enormous problem caused by *P. falciparum* in Sub-Saharan Africa. Both technological advances enabling the sequencing of the *P. vivax* genome and a recent call for worldwide malaria eradication have placed a new emphasis on the importance of addressing *P. vivax* as a major public health problem. However, because of this parasite's biology, it is especially difficult to interrupt the transmission of *P. vivax*, and experts agree that the available methods for preventing and treating both infections with *P. vivax* and *P. falciparum* are inadequate [2]. Malaria, perhaps one of the most serious and widespread diseases encountered by mankind, continues to be a major threat to about 40% of the world's population, especially in the developing world. As malaria vaccines remain problematic, chemotherapy still is the most important weapon in the fight against the disease. However, almost all available drugs have been compromised by the highly adaptable parasite, and the increasing drug resistance of *P. falciparum* continues to be the main problem. Therefore, the limited clinical repertoire of effective drugs and the emergence of multi-resistant strains substantiate the need for new proteins, or the discovery of

* Corresponding author. Tel.: +34 981 563100; fax: +34 981 594912.

E-mail addresses: humberto.gonzalez@usc.es, gonzalezdiaz@yahoo.es (H. González-Díaz).

new functions for known proteins, that may become targets of new anti-malarial compounds or the discovery of proteins involved in multi-drug resistance [3–8]. It is thus imperative that the development of new methods and strategies becomes a priority [2]. In this regard, stable protein–protein complexes formed by Protein–Protein Interactions (PPIs) may become interesting targets for new drugs and other treatment methods or strategies. For instance, there are high-molecular-weight rhoptry proteins of *P. falciparum* in a multi-protein complex consisting of proteins of 140, 130, and 110 kDa. The complex of rhoptry proteins binds to human and mouse erythrocyte membranes in association with a 120 kDa SERA protein. These proteins are believed to participate in the process of erythrocyte invasion. Sam-Yellowed have used six different antibodies (polyclonal and monoclonal) known to precipitate the high-molecular-weight rhoptry protein complex to analyze the structural relationship of proteins within the complex. The results provided insights concerning the mechanism of protein–protein interaction within the complex [9].

These types of results indicate that physically stable protein–protein biopolymer complexes (pPPC) made up of unique PPIs of *Plasmodium* sp. parasites (pPPIs) and not present in humans or other hosts may be promising targets for the development of safe drugs with low toxicity. On the contrary, the prediction of non-pPPC (non-unique *Plasmodium* sp. parasites but also present in humans) may become a source for the discovery of targets related to drug resistance not only for the treatment of malaria but also of human cancer. For instance, Human Dihydrofolate Reductase (DHFR) constitutes a primary target for antifolate drugs in cancer treatment, whereas DHFRs from *P. falciparum* and *P. vivax* are primary targets in the treatment of malaria. A recent review [10] has discussed the structural and functional impact of active-site mutations with respect to enzyme activity and antifolate resistance of DHFRs from mammals, protozoa and bacteria. DHFR is a monomeric protein with only one chain in structures deposited in PDB. However, DHFR synthase is a non-pPPC polymeric protein, which is directly involved in DHFR synthesis and consequently in drug resistance. For instance, the structure of DHFR synthase reported in the file with PDB-ID 3HBB is a PPC with four different protein chains. In this regard, a computational model able to predict non-pPPC such as DHFRs may be interesting for the prediction of protein targets involved in drug resistance in both parasite and mammalian, which may be useful in the design of chemo-protective agents.

In any case, the high number of possible genes/proteins discovered in genome/proteome of *Plasmodium* sp. determines a higher number of possible pPPC/non-pPPC structures derived from different PPIs in parasite and human hosts, which makes difficult the exhaustive experimental investigation in terms of time and resources [11,12]. In fact, many researchers in the field of Molecular and Biochemical Parasitology have recognized the high importance of different computational tools (statistical models, servers, databases) to study the proteome and/or genome of *P. falciparum* and *P. vivax* [13–18]. This fact determines that the development of predictive models for pPPIs/non-PPIs discrimination becomes a very useful tool aimed at discovering new drug targets. There are many theoretical methods for the prediction of PPIs in humans and other organisms. Many of them are based on the same approaches used for the study of protein structure–function relationships but extended to PPIs such as: sequence alignment techniques, phylogenic techniques, or alignment-free parameters besides other methods, like molecular modeling, incorporate knowledge about the 3D structure of the proteins involved in the PPIs. These methods often make use of complex trees representations (as input or output of the analysis) to represent these interactions as PPIs trees. Sequence-only methods are

often faster than 3D ones and need less structural information. On the contrary, 3D methods give a more clear idea on the structure of the protein and may be used to predict proteins with known spatial structure but unknown function [19–27]. The importance of these latter methods is that these functionally non-annotated structures become common in the Protein Data Bank (PDB) with the development of powerful characterization techniques [28]. Another role of the computational methods is the possibility to study not only the wild-type proteins but also the computational analysis of mutations [29–33]. Specifically, in this work, we are interested in computational methods to predict pPPIs that determine the formation of non-covalent but physically stable PPCs between two proteins that can be isolated and the 3D structure, chemically characterized as a potential drug target. Protein complexes are fundamental for understanding principles of cellular organizations. As the sizes of PPI trees are increasing, accurate and fast protein complex prediction from these PPI trees can be useful as a guide for biological experiments to discover novel protein complexes [34]. Otherwise, it is the direct prediction of complexes by protein–protein docking but it may become computationally expensive if we aim at performing the screening of large databases [35]. It is also of major importance to recall that nowadays it is not enough to develop a predictive model; we should also implement it into public servers, preferably of free access, for the use of the scientific community. The server packages developed by Chou and Shen [36–39], which predict the function of proteins from structural parameters or explore protein structures, are good examples in this regard. In any case, to the best of our knowledge, there is no web server available in the literature or at least a theoretical method to predict unique pPPC in *Plasmodium* and not present in humans or other parasites or hosts, based on the 3D structure of the two proteins involved in pPPIs or non-PPIs interactions.

Besides, González-Díaz et al. introduced the method called MARKovian CHEmicals IN Silico DESIGN (MARCH-INSIDE 1.0) for the computational design of small-sized drugs. In successive studies, we have extended this method to perform fast calculation of 2D and 3D alignment-free numeric parameters to describe RNA secondary structures based on molecular vibration information [40], and 3D structure of proteins based on Van der Waals [41] or electrostatic interactions [42]. Recently, the method has been renamed as MARKov CHains INVariants for NETWORKS SIMulation & DESIGN (MARCH-INSIDE 2.0). The approach uses a Markov Chain model (MCM) to calculate parameters of small-sized and also complex chemical structures [43–45]. To this end, MARCH-INSIDE describes the system as a stochastic matrix of interactions and/or transitions between the parts of the system and associates this matrix to a graph or complex network representation of this system, at the same time. This describes more adequately the broad uses of the method to numerically characterize the structure of drugs [46], RNA [40], and proteins [41,47,48], as well as drug–drug networks [49], drug–protein interactions [50], PPIs, and other systems such as an MCM associated to a graph. In this regard, MARCH-INSIDE uses networks similar to other known in proteomics, molecular, biology, and molecular microbiology, where the nodes (connected by links) are atoms (bonds), amino acids (electrostatic interactions), proteins (PPIs), genes (co-expression), organisms and microorganisms (parasite–host interactions) [51–58]. In Fig. 1 we depict the 3D structure and the Van der Waals surface for Thioredoxin (PDB-ID SYRC) a pPPC present in *P. falciparum* clone 3d7 (A) and the respective protein structure complex network graph for one of the proteins of the pPPC (B). At this structural level, the nodes are amino acids and we link two nodes with an edge if the distance between them is lower than 15 Å (this type of network is also known as contact map or protein residue networks) [59–66]. In a very recent review, we have discussed the details and many

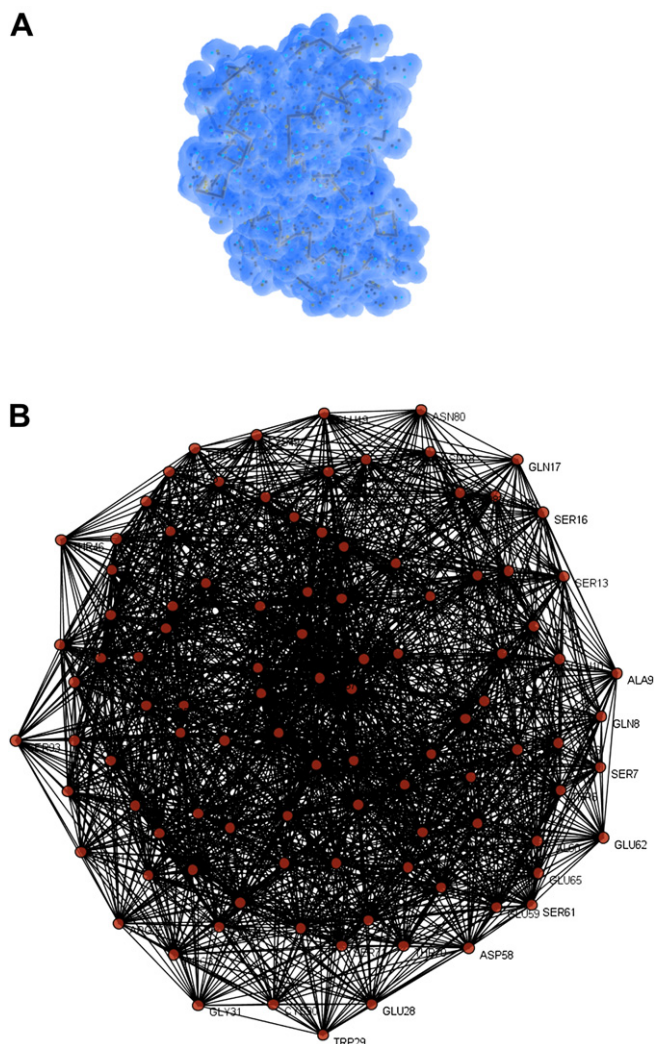


Fig. 1. 3D structure and Van der Waals surface for a *P. falciparum* protein (A) and complex network (B).

applications of the MARCH-INSIDE method to Molecular Microbiology [67].

The last upgrade of MARCH-INSIDE (carried out by Munteanu and González-Díaz) was the implementation of the Internet portal Bio-AIMS (<http://miaja.tic.udc.es/Bio-AIMS/>) with different web server packages that may be used to predict different functions of proteins from PDB files. These servers are inspired on the same philosophy of online free access and use by all the international research community, as mentioned in the previous paragraph. In particular, the server called TargetPred package offers two new Protein-QSAR servers. The first, ATCUNPred (<http://miaja.tic.udc.es/Bio-AIMS/ATCUNPred.php>) is available for prediction of ATCUN-mediated DNA-cleavage anticancer proteins [68]. The second server, EnzClassPred is available at <http://miaja.tic.udc.es/Bio-AIMS/EnzClassPred.php> and can be used to predict enzyme classes from PDB files without function annotation [69]. For all these reasons, in this work we use the MARCH-INSIDE approach for the first time to solve the problem of predicting specific pPPCs from the 3D structure of two proteins that may undergo pPPCs or not. Last but not least, we implemented the predictor in a new web server named PlasmodPPI freely available to public at <http://miaja.tic.udc.es/Bio-AIMS/PlasmodPPI.php>. In Fig. 2 we depict a flowchart for all the steps taken in this work to generate the new classifiers and server.

2. Materials and methods

2.1. Electrostatic entropy measures for PPIs

In previous works we have used different entropy invariants derived from an MCM to describe the 3D structure of one protein backbone in structure–property relationship studies. The $\theta_k(R)$ parameters used represent the average electrostatic entropy (θ) due to the interactions between all pairs of amino acids allocated inside a specific protein region (R) and placed at a distance k from each other. In this work we want to use $\theta_k(R)$ values of two proteins, $\theta_k(^1R)$ for protein 1 and $\theta_k(^2R)$ for protein 2, in order to generate structural parameters describing PPI between these proteins. To this end, we introduced herein for the first time a new type of PPI invariants in the sense that they do not depend on the interchange of proteins so that we do not need to label and distinguish them for calculation. We introduce, with this aim, three types of invariants (ti) ${}^{ti}\theta_k(R)$: PPI Average Entropy Invariant ($ti = a$), PPI Entropy Difference Invariant ($ti = d$), and PPI Entropy Product Invariant ($ti = p$):

$${}^a\theta_k(R) = {}^a\theta_k(^1R_1, ^2R_1) = \frac{1}{2}[\theta_k(^1R_1) + \theta_k(^2R_1)] \quad (1)$$

$${}^d\theta_k(R) = {}^d\theta_k(^1R_1, ^2R_1) = |\theta_k(^1R_1) - \theta_k(^2R_1)| \quad (2)$$

$${}^p\theta_k(R) = {}^p\theta_k(^1R_1, ^2R_1) = \theta_k(^1R_1) \cdot \theta_k(^2R_1) \quad (3)$$

Notably, in order to guarantee that these parameters are invariant to protein labeling as 1 or 2, we have to always use the same ${}^1R = {}^2R = R$ and $k_1 = k_2 = k$ values. In order to calculate the $\theta_k(R)$ values for each protein the method uses as a source of protein macromolecular descriptors the stochastic matrices ${}^1\Pi_e$ built up as squared matrices ($n \times n$), where n is the number of amino acids (aa) in the protein. The subscript e points to the electrostatic type of molecular force field. In previous works we have predicted the protein function based on $\theta_k(R)$ values for different types of interactions or molecular fields. The main types of molecular fields used are the following: Electrostatic, vdW, and HINT entropies. In this paper, we calculated $\theta_k(R)$ values only for Electrostatic entropies. These values have been used herein to calculate PPIs invariants and next as inputs to generate the QSAR model (see description of PPI invariants above). However, the detailed explanation for the calculation of $\theta_k(R)$ values has been published before. As follows, we give the formula for ${}^k\theta(R)$ values and some general explanations [41,67,70]:

$$\theta_k(R) = - \sum_{j=1}^n {}^kP_j(R) \cdot \log[{}^kP_j(R)] \quad (4)$$

It is remarkable that the average entropy measures depend on the absolute probabilities ${}^kP_j(R)$ according to which the amino acid j th has an electrostatic interaction with the rest of amino acids that lie within the same protein region R . These probabilities refer to amino acids placed at a distance equal to k -times the cut-off distance ($r_{ij} = k \cdot r_{\text{cut-off}}$). The method uses a Markov Chain Model (MCM) to calculate these probabilities, which also depend on the 3D interactions between all pairs of amino acids placed at distance r_{ij} in 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 displayed as a protein structure complex network. In this network the nodes are the C_α atoms of the amino acids and the edges connect

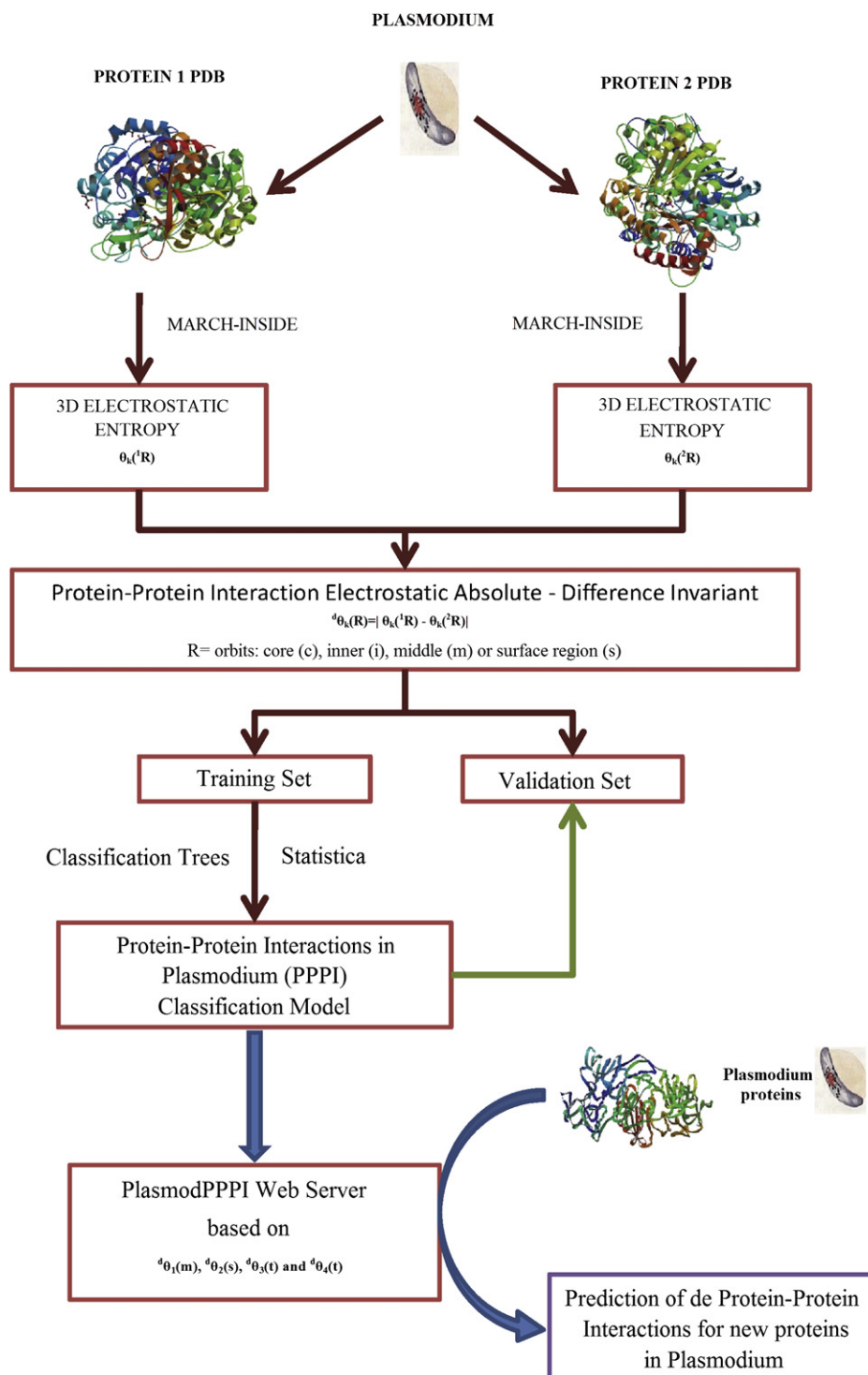


Fig. 2. Example of spatial distribution of core, inner, middle, and surface amino acids.

pairs of amino acids with $\alpha_{ij} = 1$. Euclidean 3D space $r_3 = (x, y, z)$ coordinates of the C_α atoms of amino acids listed on protein PDB files. For the calculation, all water molecules and metal ions were removed [67]. All calculations were carried out with our in-house software MARCH-INSIDE 2.0 [71].

For the calculation, the MARCH-INSIDE 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 Orbits or Regions (R). These regions are often defined in

geometric terms and called core, inner, middle or surface region. In Fig. 3 we represented the orbits of protein (*c* corresponds to core, *i* to inner, *m* to middle, and *s* to surface orbits, respectively). The diameters of the orbits, are: $0 \leq \text{orbit } c < 25$, $25 \leq \text{orbit } i < 50$, $50 \leq \text{orbit } m < 75$, and $76 \leq \text{orbit } s \leq 100$; expressed in terms of percentage of the longest distance r_{\max} with respect to the center of charge. Additionally, we take into consideration the total orbit (*t*) that contains all the amino acids in the protein (orbit diameter 0–100% of r_{\max}). Consequently, we can calculate different $\theta_k(R)$ for the

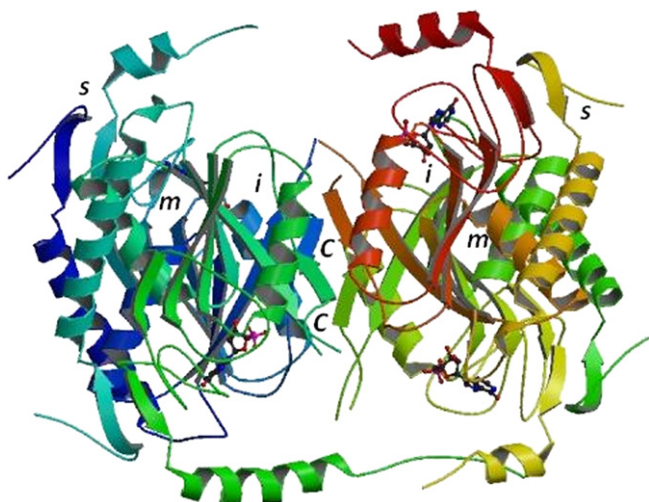


Fig. 3. Flowchart for all the steps given in the construction of the classifiers and server.

amino acids contained in an orbit (c , i , m , s , or t) and placed at a topological distance k within this orbit (k is the order named) [72–75]. In this work, we calculated altogether $5(\text{types of regions}) \times 6(\text{orders considered}) = 30 \theta_k(R)$ indices for each protein.

In order to carry out the calculations referred to in equation (1) for any kind of entropy and detailed in the previous equations, for electrostatic entropy, the elements (${}^1p_{ij}$) of ${}^1\Pi_e$ and the absolute initial probabilities ${}^A p_0(j)$ were calculated as follows [67]:

$${}^1p_{ij} = \frac{\alpha_{ij} \cdot E_{ij}}{\sum_{m=1}^{\delta+1} \alpha_{im} \cdot E_{im}} = \frac{\alpha_{ij} \cdot \frac{q_i \cdot q_j}{(d_{ij})^2}}{\sum_{m=1}^{\delta+1} \alpha_{im} \cdot \frac{q_i \cdot q_m}{(d_{im})^2}} \quad (5)$$

$${}^A p_0(j) = \frac{\frac{q_j}{d_{0j}}}{\sum_{m=1}^n \frac{q_m}{(d_{0m})^2}} \quad (6)$$

where q_i and q_j are the electronic charges for amino acids i th-aa and the j th-aa and the neighborhood relationship (truncation function $\alpha_{ij} = 1$) is turned on if these amino acids participate in a peptidic hydrogen bond or $d_{ij} < d_{\text{cut-off}} = 5 \text{ \AA}$ [67]. In this regard, the truncation of the molecular field is usually applied to simplify all the calculations in large biological systems. The distance d_{ij} is the Euclidean distance between the C_α atoms of the two amino acids and d_{0j} the distance between the amino acid and the center of charge of the protein. Both kinds of distances were derived from the x , y and z coordinates of the amino acids collected from the protein PDB files. All calculations were carried out with our in-house software MARCH-INSIDE. All water molecules and metal ions were removed for the calculation [67].

2.2. LDA models

LDA is frequently used for classification/prediction problems in physical anthropology, but it is unusual to find examples where researchers consider the statistical limitations and assumptions required for this technique. In this work, all LDA models have been trained with the software STATISTICA 6.0[®], for which our laboratory holds rights of use [76]. In LDA we use several variable selection techniques to seek the model: i) *All Effects* (include all parameters), ii) *Forward-stepwise*, iii) *Forward-entry*, iv) *Backward-stepwise*, v) *Backward-removal*, and vi) *Best subsets*. Unless we specify a different value, we always set a prior probability of

$p(\text{pPPI}) = p(\text{npPPI}) = 0.5$. The LDA discriminant equation was obtained using as input the three types of PPI invariants ${}^t\theta_k(R)$. The general form of the equation obtained by LDA is:

$$S(\text{pPPC}) = \sum_{R,k,t}^{5,5,3} a_{R,k,t} \cdot {}^t\theta_k(R) + a_0 \quad (7)$$

$S(\text{pPPC})$, the output of this model, is a real value variable that scores the propensity of a protein pair to undergo a pPPI interaction and not npPPIs forming a physically stable PPCs only in *Plasmodium* sp. The χ^2 and p -level value were examined in order to test the statistical significance of the model. The Accuracy, Specificity, Sensitivity were used to quantify the goodness-of-fit and the discriminatory power of the model. Different authors like have applied this type of LDA model using different classes of input variables to construct QSAR models for proteins or nucleic acids [77–80].

2.3. CT models

CTs have been used to test a non-linear model which is not based on assumptions of parametric distribution of data as well as non-linear models [81]. We used as Ordered Predictors the variables obtained in the Forward stepwise of the LDA. Starting from now on, several split methods were carried out: i) CT Discriminant-based Linear Combinations (CT-LC), ii) Discriminant-based univariate splits (CT-US), and CRT-style exhaustive search from univariate splits (CRT). In CRT we used three different measures of Goodness-of-fit Gini Measure, Chi-Square, and G-Square. Like in LDA we always set a prior probability of $p(\text{pPPI}) = p(\text{npPPI}) = 0.5$, unless we specify a different value. Last, we used a FACT-style direct stopping rule with a value of 0.01 to control the length of the CT. All the CTs have been trained with the software STATISTICA 6.0[®], for which our laboratory holds rights of use [76].

2.4. Dataset

The protein structures were downloaded from PDB [82] using the following schemes for PDB-database search: (i) introducing the name of the parasite species (*Plasmodium*) as input parameter in the search item called source organism (for positive cases) or (ii) introducing the PDB-IDs for all the proteins contained in the list reported in the article of Dobson and Doig [83]. The positive cases (pPPI) are those protein–protein pairs that make up a stable complex that has been structurally characterized (3D structure) in *Plasmodium* species (*Plasmodium* sp). The list of negative cases (npPPI), search scheme (b), contain enzymes and other proteins present in humans and many other organisms including other parasites that are not present in *Plasmodium* sp. The dataset consisted of 5257 pairs of proteins (774 pPPIs and 4483 npPPIs) from more than 20 organisms, including parasites and human or cattle hosts. Altogether, 581 pPPIs and 3395 npPPIs were used in training and 193 pPPIs and 1088 npPPIs were used in validation. Detailed information about the PDB-ID, the values of the electrostatic entropy indices, the corresponding observed classification, and the predicted classification for each pPPI or npPPI pair are given in the [Supporting information](#).

3. Results and discussion

Several researchers have demonstrated the high performance of different types of computational classifiers in protein or PPI structure–function relationship studies based on different algorithms as is the case, for instance, of the works carried out by Chou

et al. [84–90], Fernandez and Caballero [91–93]. In particular, the LDA algorithm, a simpler type of the classifier used herein, was employed to train linear models based on different combinations of parameters [94].

3.1. Linear discriminant analysis (LDA) models

A simple Linear Discriminant Analysis (LDA), with only four variables, was developed to assign each protein pair as pPPI or npPPI. The best equation found was:

$$S(\text{pPPC}) = -0.09506 \cdot {}^d\theta_3(m) - 0.02219 \cdot {}^d\theta_4(s) \\ - 0.62697 \cdot {}^d\theta_5(t) + 0.51126 \cdot {}^d\theta(t) - 0.30646 \\ N = 3976 \quad \chi^2 = 947.95 \quad p < 0.00 \quad (8)$$

The statistical parameters for the above equation are: Number of protein entries in training (N), Chi-square statistic (χ^2), and error level (p -level), which have to be <0.05 [95]. All the statistical data of this model are summed up in Table 1. The discriminant function reported in the results section presented statistically significant results of goodness-of-fit for both training and validation series, carried out with an external series of pPPI and npPPI that were never used to train the model. Interestingly four variables, ${}^d\theta_3(m)$, ${}^d\theta_4(s)$, ${}^d\theta_5(t)$ and ${}^d\theta(t)$, out of more than 30 parameters calculated appear in many models. These parameters have the general formula ${}^d\theta_k(R) = |\theta_k(R)_{\text{prot1}} - \theta_k(R)_{\text{prot2}}|$, which are the absolute difference between the electrostatic entropy values $\theta_k(R)$ for amino

acids on the surface of the two proteins forming the PPI pairs. This fact indicates that the difference between the surface electrostatic entropy is very important not only for PPI interactions in general but also to discriminate the unique complex present in *Plasmodium* sp. (pPPIs) and not in other organisms. The model presents a good overall classification of pPPI and npPPI. This level of accuracy is generally accepted by other researchers that have applied LDA to find QSAR models useful in molecular parasitology and related areas; e.g., the works of García-Domenech, Marrero-Ponce, Bruno-Blanch, Galvez, Gozalbes and others predicting active compounds against *Trypanosoma cruzi*, *Mycobacterium avium*, *Toxoplasma gondii*, *P. falciparum*, *Trichomonas vaginalis*, *Fasciola hepatica*, and other parasites [96–100]; see also the works of Marrero-Ponce on protein and DNA/RNA QSAR studies [101–103].

3.2. Artificial neural network (ANN) models

The comparison of linear and non-linear models is essential to test how directly our parameters are correlated to the biological property [104]. The automatic selection of variables (features) was activated for all models. In particular, the Linear Neural Network (LNN) algorithm and other types of Artificial Neural Network (ANN), were used herein to train different linear and non-linear models based on different combinations of parameters. Table 1 also depicts the results for the best models found. The profile of the ANN model was specified with a simple notation as follows: ANN type N_{iv} : N_{in} - N_{H1} - N_{H2} - N_{on} : N_{ov} . The ANN types presented, besides LNN, are Multi-Layer Perceptron (MLP), Probabilistic Neural Network (PNN), and Radial Basis Function (RBF) [105]. The parameter N_{iv} is the number of input variables, N_{in} is the number of input neurons (one per input variable), N_{H1} is the number of neurons in the first Hidden layer (H1), N_{H2} is the number of neurons in the second Hidden layer (H1), N_{on} is the number of output neurons, and N_{ov} is the number of output variables.

Table 1
Summary of results for LDA, CT, and ANN analysis.

Technique			Training sub-set			Validation sub-set		
Profile	Parameters	Group	%	npPPI	pPPI	%	npPPI	pPPI
LDA	Specificity	npPPI	85.0	2886	509	82.4	897	191
	Sensitivity	pPPI	94.8	30	551	92.7	14	179
	Accuracy	Total	86.4	–	–	84.0	–	–
CT	Specificity	npPPI	98.5	3343	52	98.0	1066	22
	Sensitivity	pPPI	91.2	51	530	90.2	19	174
	Accuracy	Total	97.4	–	–	96.8	–	–
US	Specificity	npPPI	95.6	3247	148	96.5	1050	38
	Sensitivity	pPPI	83.8	94	487	84.5	30	163
	Accuracy	Total	93.9	–	–	94.7	–	–
CRT	Specificity	npPPI	97.6	3315	80	97.8	1064	24
	Sensitivity	pPPI	84.7	89	492	83.4	32	161
	Accuracy	Total	95.7	–	–	95.6	–	–
Chi-square	Specificity	npPPI	97.6	3315	80	97.8	1064	24
	Sensitivity	pPPI	84.7	89	492	83.4	32	161
	Accuracy	Total	95.7	–	–	95.6	–	–
G-square	Specificity	npPPI	98.6	3348	47	98.4	1071	17
	Sensitivity	pPPI	81.8	106	475	80.3	38	155
	Accuracy	Total	96.2	–	–	95.7	–	–
MLP	Sensitivity	pPPI	83.3	484	97	82.9	160	33
	Specificity	npPPI	84.0	544	2851	82.9	186	902
	Accuracy	Total	83.9	–	–	82.9	–	–
MLP	Sensitivity	pPPI	83.1	483	98	81.9	158	35
	Specificity	npPPI	83.0	577	2818	81.6	200	888
	Accuracy	Total	83.0	–	–	81.7	–	–
RBF	Sensitivity	pPPI	18.9	110	471	20.2	39	154
	Specificity	npPPI	17.3	2807	588	15.5	919	169
	Accuracy	Total	17.6	–	–	16.2	–	–
LNN	Sensitivity	pPPI	92.6	538	43	90.2	174	19
	Specificity	npPPI	92.2	264	3131	90.4	104	984
	Accuracy	Total	92.3	–	–	90.4	–	–

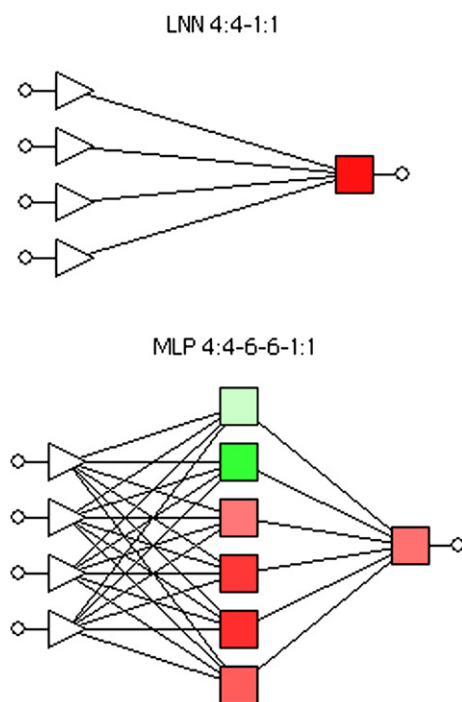


Fig. 4. Illustrative example of the topology used for different ANNs trained in this work.

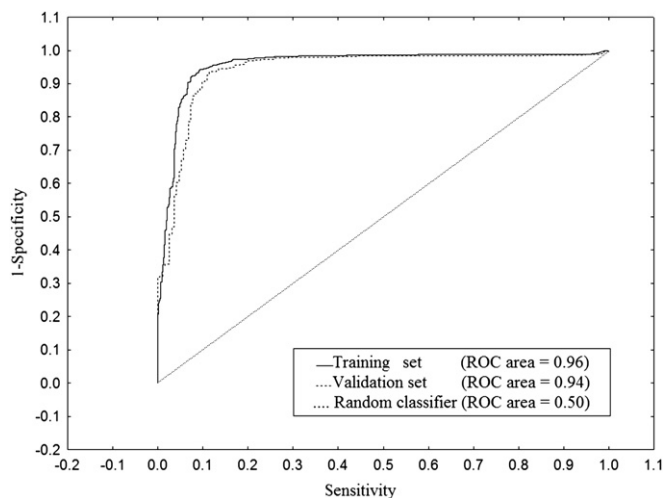


Fig. 5. ROC curve for pPPC predictor.

In particular, the model LNN 4:4–1:1 is the simplest model found with the highest levels of Sensitivity = 92.6, Specificity = 92.2 and Accuracy = 92.3 in the training set. These values are excellent considering that this predictor uses only two molecular descriptors of the PPI pair, which is a very complex structure in chemical terms, to fit a large data set of 581 pPPIs and 3395 npPPIs. The profile 4:4–1:1 indicates that this model assign the values of four input variables to four input neurons that perform a weighed sum and assigns the result to one output neuron; which gives the final result of the case classification according to the threshold value that has been optimized. In addition, the model LNN 4:4–1:1 also presented a higher levels of Sensitivity = 90.2, Specificity = 90.4 and Accuracy = 90.4 in external validation (test) set (see Table 1). In Fig. 4 we illustrate the topology of this LNN network compared with a non-linear ANN. Interestingly, four variables

$d\theta_3(m)$, $d\theta_4(s)$, $d\theta_4(t)$ and $d\theta_5(t)$, out of more than 30 parameters calculated, appear in many models. This fact indicates that the difference between the electrostatic entropy is very important not only for PPI interactions in general but also to discriminate a unique complex present in *Plasmodium* (pPPIs). On the other hand, the product and average invariant types (${}^a\theta_k(R)$ and ${}^p\theta_k(R)$) do not appear to be relevant.

We also validated the linear model by means of a ROC curve analysis (see Fig. 5) to demonstrate that there is a linear and not an indirect non-linear relationship between our indices and the classification of pPPCs [106]. The values of the area under the ROC curve for this model are 0.95 and 0.96 very close to 1 (the highest possible value) and notably different from 0.5 (the typical value of a random classifier). This kind of analysis is an accepted tool in Bioinformatics to demonstrate which classification methods outperform the other methods, e.g. the study carried out by Xu and Du related to PPIs [107] or the work of Mahdavi and Lin [108]. This first search points to a linear instead of non-linear relationship between pPPI prediction and $d\theta_k(R)$ values, giving additional proofs of the validity of our methodology. For instance, in Table 1 we can see that more complicated models with non-linear profiles do not improve the linear model and give even worse results sometimes.

3.3. Classification Tree (CT) models

Last, considering that non-linear ANN did not notably improved LDA, we used the variables pre-selected by LDA as inputs for a Classification Tree (CT) analysis. With complete data sets, LDA may be a simpler and sometimes better choice. However, the testing of data prior to analysis is necessary, and CTs are recommended either as a replacement for LDA or as a supplement whenever data do not meet relevant assumptions [109]. Table 1 also depicts the results for the best CT models found. The automatic selection of variables (features) was activated for all models if available. In Fig. 6 we illustrate the graph representation

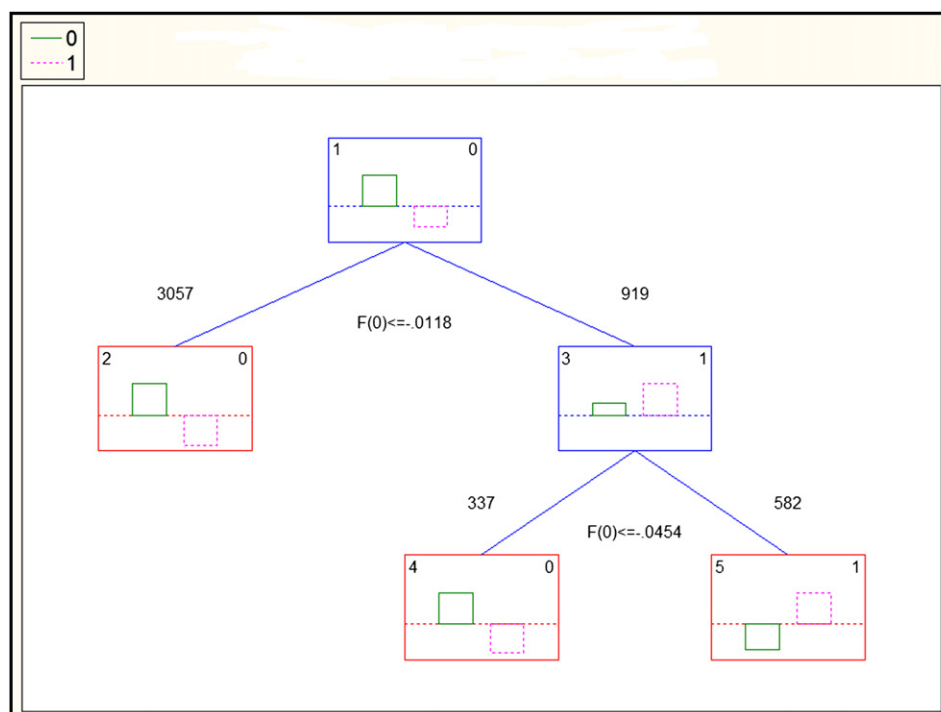


Fig. 6. Structure of the CT model found.


Table 2
Structure of the CT-LC model.

Parameters	Parent nodes				
	1	2	3	4	5
Child nodes					
Left branch	2		4		
Right branch	3		5		
npPPI	3395	3018	377	325	52
pPPI	581	39	542	12	530
Predicted class	npPPI	npPPI	pPPI	npPPI	pPPI
Split conditions ($LC_i \leq$ Split constant)					
LC_i	LC_1	LC_2	LC_3	LC_4	LC_5
Split constant	0.011758	0	0.045360	0	0
$d\theta_3(m)$	-0.000827	0	-0.004075	0	0
$d\theta_4(s)$	-0.000193	0	-0.001044	0	0
$d\theta_4(t)$	0.005454	0	0.018150	0	0
$d\theta_5(t)$	-0.004447	0	-0.014544	0	0

of the CT-LC trained in this work and in Table 2 we give details about the structure of this CT and the split rules derived. In particular, the model CT-LC is the simplest CT model found with the highest levels of Sensitivity = 91.2% Specificity = 98.5% and Accuracy = 97.4% in the training set. These values are excellent considering that this predictor uses only two molecular descriptors of the PPI pair; which is a very complex structure in chemical terms, to fit a large data set of 582 TPPIs and 3394 non-TPPIs (see Table 1). In fact, the CT analysis yielded the best model found in this work.

3.4. PlasmodPPI, a server for PPC plasmodium targets


Last, we have to consider that with the advent of Internet it is important not only to develop new predictive models for proteome research but also to carry out the implementation of these models in public web servers available to other research groups [36–39,110–113]. In this regard, we implemented this predictor into a web server freely available to public at <http://miaja.tic.udc.es/Bio-AIMS/PlasmodPPI.php>. This is the first model and web server that



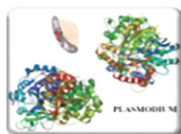
RGB Groups
RNASA, TIC
University of A Coruña
Microbiology & Parasitology
University of Santiago de Compostela
Spain

PlasmodPPI @ Bio-AIMS

Modelling the reality



Home | Theory | About



Plasmod-PPI
Plasmodium Protein-Protein Interactions (PPPI)


Tool: MARCH-INSIDE (Python version)
Data: RCSB PDB

PDB-chain lists : Please paste the names of the PDB chains as two lists (max. 50)

Notes: There is no space between the PDB name and the chain label, no empty new line; the results will print the pairs between the chain from list 1 with the chain from list 2 (not the combination of the list items).

3C5IA
2F6IE
1SYRC


3C5IE
2GHUA
1SYRF



RGB Groups
RNASA, TIC
University of A Coruña
Microbiology & Parasitology
University of Santiago de Compostela
Spain

PlasmodPPI.calc @ Bio-AIMS

Modelling the reality



Home | Theory | About

Process ID = 137494af1487ed0412
PDB List 1 = 3C5IA 2F6IE 1SYRC
PDB List 2 = 3C5IE 2GHUA 1SYRF

... please wait

PDB Update/Verification [List 1] ...
3C5IA 2F6IE 1SYRC

PDB Update/Verification [List 2] ...
3C5IE 2GHUA 1SYRF

Processing PDB-chain List 1 ...
3C5IA 2F6IE 1SYRC
Processing PDB-chain List 2 ...
3C5IE 2GHUA 1SYRF

Result file = Results/137494af1487ed0412/PlasmodPPI.calc.txt

Calculated at 2009-11-04 10:25:19

Chain1	Chain2	Complex
3C5IA	3C5IE	YES
2F6IE	2GHUA	NO
1SYRC	1SYRF	YES

Fig. 7. Example of use of PlasmodPPI web tool: (A) Input and (B) Output pages.

predicts how unique is a protein–protein complex in Plasmodium proteome with respect to other parasites and hosts breaking new ground for anti-plasmodium drug target discovery.

In order to demonstrate the practical utility of this Web server, three examples of protein chain pairs have been used to evaluate the possibility to make up unique complexes in Plasmodium, a human pathogen parasite: 3C5IA–3C5IE, 2F6IE–2GHUA and 1SYRC–1SYRF. Fig. 7 presents the input (A) and output (B) web pages of the PlasmodPPI tool. The first pair contains the first chain A of the *Plasmodium knowlesi* choline kinase (a transferase, 3C5I) and the cleaved fragment of N-terminal expression tag (chain E), all expressed in *Escherichia coli*. Choline kinase is the first enzyme in the Kennedy pathway (CDP-choline pathway) for the biosynthesis of the most essential phospholipid, phosphatidylcholine, in Plasmodium. In addition, choline kinase also plays a pivotal role in trapping essential polar head group choline inside the malaria parasite. The inhibition of choline kinase will lead to a decrease in phosphocholine, which in turn causes a decrease in phosphatidylcholine biosynthesis, resulting in death of the parasite. This pair of protein chains is evaluated to make up the unique complex in Plasmodium that can be a target for new anti-parasite drugs. The second pair example is formed by the chain E of the 2F6I hydrolase [114], a ATP-dependent CLP protease (serine-type endopeptidase) from *Plasmodium falciparum* (expressed in *E. coli*) and the chain A of the 2GHU hydrolase, Falcipain-2 (FP-2) of *P. falciparum* [115]. FP-2 is a papain-family (C1A) cysteine protease that plays an important role in the parasite life cycle by degrading erythrocyte proteins, most notably hemoglobin. Inhibition of FP-2 and its paralogues prevents parasite maturation. These two chains of hydrolases are not evaluated by our tool to form a unique complex. This can be explained by the different targets of these hydrolases and different cellular localizations (2F6I in cytoplasm and 2GHU in food vacuole for hemoglobin degradation and cleavage of cytoskeletal elements). The last example is formed by the chains C and F of the 1SYR protein, a *Plasmodium falciparum* thioredoxin in the genetic structure with an unknown function [116]. These chains are evaluated to form a unique complex according to the localization of both chains in the same protein. PlasmodPPI tool can become important for the discovery of new anti-plasmodium drug targets and can be useful as model for building similar models for other types of parasites or other organisms.

4. Conclusions

The overall findings suggest that the new type of parameters introduced herein is useful to numerically characterize the structure of PPCs, formed after PPIs, in protein structure–function studies. We also demonstrate that it is possible to distinguish between PPCs (pPPCs cases) formed according to unique PPIs in *Plasmodium* sp. (pPPIs) and not present in other parasites or host organisms using these parameters. We generate and compare linear and non-linear classifiers. We show that it is possible to predict PPIs that undergo pPPC formation with a simple linear classifier based on the absolute difference between 3D protein surface electrostatic entropies of the pair proteins. The model was implemented in a public web server, available for free-of-charge use to the scientific community.

Acknowledgments

We thank the kind and professional attention of Prof. J.E. Mark (Computational & Theoretical Polymer Science editor for Polymer) as well as the opinion of the reviewers. Gonzalez–Díaz H. and Munteanu C.R. acknowledge research contract financed by the Contract/grant sponsor: Isidro Parga Pondal Program, Xunta de

Galicia. The authors thank for the partial financial support from the grants 2007/127 and 2007/144 from the General Directorate of Scientific and Technological Promotion of the Galician University System of the Xunta de Galicia and from grant (Ref. PIO52048 and RD07/0067/0005) funded by the Carlos III Health Institute.

Appendix. Supplementary data

Supplementary data associated with this article can be found in online version, at doi:10.1016/j.polymer.2009.11.029.

References

- [1] Verra F, Mangano VD, Modiano D. Parasite Immunol 2009;31(5):234–53.
- [2] Mueller I, Galinski MR, Baird JK, Carlton JM, Kochar DK, Alonso PL, et al. Lancet Infect Dis 2009;9(9):555–66.
- [3] Bonilla JA, Bonilla TD, Yowell CA, Fujioka H, Dame JB. Mol Microbiol 2007;65(1):64–75.
- [4] Turschner S, Efferth T. Mini Rev Med Chem 2009;9(2):206–2124.
- [5] Sanchez CP, Rotmann A, Stein WD, Lanzer M. Mol Microbiol 2008;70(4):786–98.
- [6] Sanchez CP, Rohrbach P, McLean JE, Fidock DA, Stein WD, Lanzer M. Mol Microbiol 2007;64(2):407–20.
- [7] Nunes MC, Goldring JP, Doerig C, Scherf A. Mol Microbiol 2007;63(2):391–403.
- [8] Siden-Kiamos I, Ecker A, Nyback S, Louis C, Sinden RE, Billker O. Mol Microbiol 2006;60(6):1355–63.
- [9] Sam-Yellowe TY. Exp Parasitol 1993;77(2):179–94.
- [10] Volpato JP, Pelletier JN. Drug Resist Updat 2009;12(1–2):28–41.
- [11] Carucci DJ, Yates 3rd JR, Florens L. Int J Parasitol 2002;32(13):1539–42.
- [12] Coppel RL, Black CG. Int J Parasitol 2005;35(5):465–79.
- [13] Bender A, van Dooren GG, Ralph SA, McFadden GI, Schneider G. Mol Biochem Parasitol 2003;132(2):59–66.
- [14] Carlton JM, Muller R, Yowell CA, Fluegge MR, Sturrock KA, Pritt JR, et al. Mol Biochem Parasitol 2001;118(2):201–10.
- [15] Coppel RL. Mol Biochem Parasitol 2001;118(2):139–45.
- [16] Cui L, Fan Q, Hu Y, Karamycheva SA, Quackenbush J, Khuntirat B, et al. Mol Biochem Parasitol 2005;144(1):1–9.
- [17] Gunasekera AM, Patankar S, Schug J, Eisen G, Kissinger J, Roos D, et al. Mol Biochem Parasitol 2004;136(1):35–42.
- [18] Huestis R, Fischer K. Mol Biochem Parasitol 2001;118(2):187–99.
- [19] Sharon I, Davis JV, Yona G. Methods Mol Biol 2009;541:61–88.
- [20] Liu L, Cai Y, Lu W, Feng K, Peng C, Niu B. Biochem Biophys Res Commun 2009;380(2):318–22.
- [21] Skrabanek L, Saini HK, Bader GD, Enright AJ. Mol Biotechnol 2008;38(1):1–17.
- [22] Najafabadi HS, Salavati R. Genome Biol 2008;9(5):R87.
- [23] Kim S, Shin SY, Lee IH, Kim SJ, Srimam R, Zhang BT. Nucleic Acids Res 2008;36(Web Server issue):W411–5.
- [24] Jaeger S, Gaudan S, Leser U, Rebholz-Schuhmann D. BMC Bioinformatics 2008;8(9 Suppl):S2.
- [25] Burger L, van Nimwegen E. Mol Syst Biol 2008;4:165.
- [26] Scott MS, Barton GJ. BMC Bioinformatics 2007;8:239.
- [27] Zvebil MJ, Tang L, Cookson E, Selkirk ME, Thornton JM. Mol Biochem Parasitol 1993;58(1):145–53.
- [28] von Grotthuss M, Plewczynski D, Ginalski K, Rychlewski L, Shakhnovich EI. BMC Bioinformatics 2006;7:53.
- [29] Lappalainen I, Thusberg J, Shen B, Vihinen M. Proteins 2008;72(2):779–92.
- [30] Shen B, Bai J, Vihinen M. Protein Eng Des Sel 2008;21(1):37–44.
- [31] Shen B, Vihinen M. Protein Eng Des Sel 2004;17(3):267–76.
- [32] Liu ML, Shen BW, Nakaya S, Pratt KP, Fujikawa K, Davie EW, et al. Blood 2000;96(3):979–87.
- [33] Shen B, Nolan JP, Sklar LA, Park MS. Nucleic Acids Res 1997;25(16):3332–8.
- [34] Chua HN, Ning K, Sung WK, Leong HW, Wong L. J Bioinform Comput Biol 2008;6(3):435–66.
- [35] Smith GR, Sternberg MJ. Curr Opin Struct Biol 2002;12(1):28–35.
- [36] Shen HB, Chou KC. Anal Biochem 2008;373(2):386–8.
- [37] Shen HB, Chou KC. Protein Eng Des Sel 2007;20(11):561–7.
- [38] Chou KC, Shen HB. Biochem Biophys Res Commun 2007; doi:10.1016/j.bbrc.2007.10.06.1027.
- [39] Chou KC, Shen HB. Nat Protoc 2008;3(2):153–62.
- [40] González-Díaz H, de Armas RR, Molina R. Bioinformatics 2003;19(16):2079–87.
- [41] González-Díaz H, Saiz-Urra L, Molina R, Santana L, Uriarte E. J Proteome Res 2007;6(2):904–8.
- [42] Gonzalez-Diaz H, Molina R, Uriarte E. FEBS Lett 2005;579(20):4297–301.
- [43] Concu R, Podda G, Uriarte E, Gonzalez-Diaz H. J Comput Chem 2009;30:1510–20.
- [44] Gonzalez-Diaz H, Saiz-Urra L, Molina R, Gonzalez-Diaz Y, Sanchez-Gonzalez A. J Comput Chem 2007;28(6):1042–8.
- [45] González-Díaz H, Pérez-Castillo Y, Podda G, Uriarte E. J Comput Chem 2007;28:1990–5.

- [46] Santana L, Uriarte E, González-Díaz H, Zagotto G, Soto-Otero R, Mendez-Alvarez E. *J Med Chem* 2006;49(3):1149–56.
- [47] Agüero-Chapin G, Varona-Santos J, de la Riva GA, Antunes A, Gonzalez-Villa T, Uriarte E, et al. *J Proteome Res* 2009;8(4):2122–8.
- [48] Concu R, Dea-Ayuela MA, Perez-Montoto LG, Bolas-Fernandez F, Prado-Prado FJ, Podda G, et al. *J Proteome Res* 2009;8(9):4372–82.
- [49] Santana L, Gonzalez-Diaz H, Quezada E, Uriarte E, Yanez M, Vina D, et al. *J Med Chem* 2008;51(21):6740–51.
- [50] Vina D, Uriarte E, Orallo F, Gonzalez-Diaz H. *Mol Pharmacol* 2009;6(3):825–35.
- [51] Bornholdt S, Schuster HG. *Handbook of graphs and complex networks: from the genome to the internet*. Weinheim: WILEY-VCH GmbH & CO. KGa; 2003.
- [52] Mazurie A, Bonchev D, Schwikowski B, Buck GA. *Bioinformatics* 2008;24(22):2579–85.
- [53] Managbanag JR, Witten TM, Bonchev D, Fox LA, Tsuchiya M, Kennedy BK, et al. *PLoS One* 2008;3(11):e3802.
- [54] Witten TM, Bonchev D. *Chem Biodivers* 2007;4(11):2639–55.
- [55] Bonchev D, Buck GA. *J Chem Inf Model* 2007;47(3):909–17.
- [56] Bonchev D. *SAR QSAR Environ Res* 2003;14(3):199–214.
- [57] Estrada E. *J Proteome Res* 2006;5(9):2177–84.
- [58] Estrada E. *Proteomics* 2006;6(1):35–40.
- [59] Gupta N, Mangal N, Biswas S. *Proteins* 2005;59(2):196–204.
- [60] Webber Jr CL, Giuliani A, Zbilut JP, Colosimo A. *Proteins* 2001;44(3):292–303.
- [61] Gobel U, Sander C, Schneider R, Valencia A. *Proteins* 1994;18(4):309–17.
- [62] Krishnan A, Zbilut JP, Tomita M, Giuliani A. *Curr Protein Pept Sci* 2008;9(1):28–38.
- [63] Krishnan A, Giuliani A, Zbilut JP, Tomita M. *PLoS One* 2008;3(5):e2149.
- [64] Palumbo MC, Colosimo A, Giuliani A, Farina L. *FEBS Lett* 2007;581(13):2485–9.
- [65] Krishnan A, Giuliani A, Zbilut JP, Tomita M. *J Proteome Res* 2007;6(10):3924–34.
- [66] Krishnan A, Giuliani A, Tomita M. *PLoS ONE* 2007;2(6):e562.
- [67] González-Díaz H, González-Díaz Y, Santana L, Ubeira FM, Uriarte E. *Proteomics* 2008;8:750–78.
- [68] Munteanu CR, Vázquez JM, Dorado J, Pazos-Sierra A, Sánchez-González A, Prado-Prado FJ, et al. *Proteome Res* 2009; doi:10.1021/pr900556g.
- [69] Concu R, Dea-Ayuela MA, Perez-Montoto LG, Prado-Prado FJ, Uriarte E, Bolas-Fernandez F, et al. *Biochim Biophys Acta* 2009; doi:10.1016/j.bbapap.2009.1008.1020.
- [70] González-Díaz H, Molina R, Uriarte E. *Bioorg Med Chem Lett* 2004;14(18):4691–5.
- [71] Gonzalez-Diaz H, Prado-Prado F, Ubeira FM. *Curr Top Med Chem* 2008;8(18):1676–90.
- [72] González-Díaz H, Saíz-Urra L, Molina R, Uriarte E. *Polymer* 2005;46(8):2791–8.
- [73] González-Díaz H, Molina-Ruiz R, and Hernandez I. MARCH- INSIDE v3.0 (MAR kov CH ains IN variants for SI mulation & DE sign); Windows supported version under request to the main author contact email: gonzalezdiaz@yahoo.es; 2007.
- [74] Cruz-Monteagudo M, Gonzalez-Diaz H. *Eur J Med Chem* 2005;40(10):1030–41.
- [75] Gonzalez-Diaz H, Agüero-Chapin G, Varona J, Molina R, Delogu G, Santana L, et al. *J Comput Chem* 2007;28(6):1049–56.
- [76] StatSoft.Inc. STATISTICA (data analysis software system), version 6.0, www.statsoft.com. Statsoft, Inc; 2002.
- [77] Marrero-Ponce Y, Medina-Marrero R, Castro AE, Ramos de Armas R, González-Díaz H, Romero-Zaldivar V, et al. *Molecules* 2004;9:1124–47.
- [78] Ramos de Armas R, Gonzalez Diaz H, Molina R, Uriarte E. *Proteins* 2004;56(4):715–23.
- [79] Ramos de Armas R, González-Díaz H, Molina R, Perez Gonzalez M, Uriarte E. *Bioorg Med Chem* 2004;12(18):4815–22.
- [80] Ramos de Armas R, González-Díaz H, Molina R, Uriarte E. *Biopolymers* 2005;77(5):247–56.
- [81] Hill T, Lewicki P. *Statistics methods and applications. A comprehensive reference for science, industry and data mining*. Tulsa: StatSoft; 2006.
- [82] Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA. *Nucleic Acids Res* 2005;33(Database issue):D183–7.
- [83] Dobson PD, Doig AJ. *J Mol Biol* 2003;330(4):771–83.
- [84] Chou KC. *J Proteome Res* 2005;4(4):1413–8.
- [85] Chou KC, Elrod DW. *J Proteome Res* 2003;2(2):183–90.
- [86] Chou KC, Shen HB. *J Proteome Res* 2006;5:1888–97.
- [87] Chou KC, Shen HB. *J Proteome Res* 2006;5:3420–8.
- [88] Chou KC, Shen HB. *J Proteome Res* 2007;6:1728–34.
- [89] Chou KC, Cai YD. *J Proteome Res* 2006;5(2):316–22.
- [90] Chou KC, Elrod DW. *J Proteome Res* 2002;1(5):429–33.
- [91] Fernández M, Caballero F, Fernández L, Abreu JI, Acosta G. *Proteins* 2008;70(1):167–75.
- [92] Caballero J, Fernandez M. *Curr Top Med Chem* 2008;8(18):1580–605.
- [93] Fernández L, Caballero J, Abreu JI, Fernández M. *Proteins* 2007;67:834–52.
- [94] Guha R, Jurs PC. *J Chem Inf Comput Sci* 2004;44(6):2179–89.
- [95] Van Waterbeemd H. Discriminant analysis for activity prediction. In: Van Waterbeemd H, editor. *Chemometric methods in molecular design*, vol. 2. New York, NY: Wiley-VCH; 1995. p. 265–82.
- [96] Garcia-Garcia A, Galvez J, de Julian-Ortiz JV, Garcia-Domenech R, Munoz C, Guna R, et al. *J Biomol Screen* 2005;10(3):206–14.
- [97] Garcia-Garcia A, Galvez J, de Julian-Ortiz JV, Garcia-Domenech R, Munoz C, Guna R, et al. *J Antimicrob Chemother* 2004;53(1):65–73.
- [98] Gozalbes R, Brun-Pascaud M, Garcia-Domenech R, Galvez J, Pierre-Marie G, Jean-Pierre D, et al. *Antimicrobial Agents Chemother* 2000;44(10):2771–6.
- [99] Gozalbes R, Galvez J, Garcia-Domenech R, Derouin F. *SAR QSAR Environ Res* 1999;10(1):47–60.
- [100] Marrero-Ponce Y, Meneses-Marcel A, Rivera-Borroto OM, Garcia-Domenech R, De Julian-Ortiz JV, Montero A, et al. *J Comput Aided Mol Des* 2008;22(8):523–40.
- [101] Marrero-Ponce Y, Ortega-Broche SE, Diaz YE, Alvarado YJ, Cubillan N, Cardoso GC, et al. *J Theor Biol* 2009;259(2):229–41.
- [102] Marrero-Ponce Y, Castillo Garit JA, Nodarse D. *Bioorg Med Chem* 2005;13(10):3397–404.
- [103] Marrero-Ponce Y. *J Chem Inf Comput Sci* 2004;44(6):2010–26.
- [104] Fernandez M, Caballero J, Tundidor-Camba A. *Bioorg Med Chem* 2006;14(12):4137–50.
- [105] Rabow AA, Scheraga HA. *J Mol Biol* 1993;232(4):1157–68.
- [106] Hill T, Lewicki P. *Statistics methods and applications*. Tulsa: StatSoft; 2006.
- [107] Xu T, Du L, Zhou Y. *BMC Bioinformatics* 2008;9:472.
- [108] Mahdavi MA, Lin YH. *Genomics Proteomics Bioinformatics* 2007;5(3–4):177–86.
- [109] Feldesman MR. *Am J Phys Anthropol* 2002;119(3):257–75.
- [110] Schlessinger A, Yachdav G, Rost B. *Bioinformatics* 2006;22(7):891–3.
- [111] Mewes HW, Frishman D, Mayer KF, Munsterkötter M, Noubibou O, Pagel P, et al. *Nucleic Acids Res* 2006;34(Database issue):D169–172.
- [112] Xie D, Li A, Wang M, Fan Z, Feng H. *Nucleic Acids Res* 2005;33(Web Server issue):W105–110.
- [113] McDermott J, Guerin M, Frazier Z, Chang AN, Samudrala R. *Nucleic Acids Res* 2005;33(Web Server issue):W324–5.
- [114] Vedadi M, Lew J, Artz J, Amani M, Zhao Y, Dong A, et al. *Mol Biochem Parasitol* 2007;151(1):100–10.
- [115] Hogg T, Nagarajan K, Herzberg S, Chen L, Shen X, Jiang H, et al. *J Biol Chem* 2006;281(35):25425–37.
- [116] Banerjee AK, Arora N, Murty US. *J Vector Borne Dis* 2009;46(3):171–83.

Complex Network Spectral Moments for ATCUN Motif DNA Cleavage: First Predictive Study on Proteins of Human Pathogen Parasites

Cristian R. Munteanu,^{*,†} José M. Vázquez,[†] Julián Dorado,[†] Alejandro Pazos Sierra,[†]
 Ángeles Sánchez-González,[‡] Francisco J. Prado-Prado,[§] and Humberto González-Díaz^{*,§}

Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruña, Campus de Elviña, s/n 15071 A Coruña, Spain, Department of Inorganic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, Praza Seminario de Estudos Galegos, s/n. Campus sur, 15782 Santiago de Compostela, Spain, and Department of Microbiology & Parasitology, Faculty of Pharmacy, University of Santiago de Compostela, Praza Seminario de Estudos Galegos, s/n. Campus sur, 15782 Santiago de Compostela, Spain

Received June 25, 2009

The development of methods that can predict the metal-mediated biological activity based only on the 3D structure of metal-unbound proteins has become a goal of major importance. This work is dedicated to the amino terminal Cu(II)- and Ni(II)-binding (ATCUN) motifs that participate in the DNA cleavage and have antitumor activity. We have calculated herein, for the first time, the 3D electrostatic spectral moments for 415 different proteins, including 133 potential ATCUN antitumor proteins. Using these parameters as input for Linear Discriminant Analysis, we have found a model that discriminates between ATCUN-DNA cleavage proteins and nonactive proteins with 91.32% Accuracy (379 out of 415 of proteins including both training and external validation series). Finally, the model has predicted for the first time the DNA cleavage function of proteins from the pathogen parasites. We have predicted possible ATCUN-like proteins with a probability higher than 99% in nine parasite families such as Trypanosoma, Plasmodium, Leishmania, or Toxoplasma. The distribution by biological function of the ATCUN proteins predicted has been the following: oxidoreductases 70.5%, signaling proteins 62.5%, lyases 58.2%, membrane proteins 45.5%, ligases 44.4%, hydrolases 41.3%, transferases 39.2%, cell adhesion proteins 34.5%, metal binders 33.5%, translation proteins 25.0%, transporters 16.7%, structural proteins 9.1%, and isomerases 8.2%. The model is implemented at <http://miaja.tic.udc.es/Bio-AIMS/ATCUNPred.php>.

Keywords: Cu–Ni cluster • ATCUN-like motif • DNA cleavage • antitumor activity • Markov model • QSAR • electrostatic potential • Plasmodium • Fasciola • Leishmania

Introduction

An important goal in bioinorganic chemistry is to find the function of a protein from the experimentally determined structure with minimum costs. Thus, the chemical databases contain numerous 3D metal-binding protein structures without any information about their biological function that depends on the metal ion type. Inside these proteins, there can be found specific amino acid sequences with high affinity for different metals.

The amino terminal Cu(II)- and Ni(II)-binding (ATCUN) motif is a small metal-binding site and was discovered for the first time in serum albumin.¹ It was proven to have antitumor activity by participating to the DNA cleavage with the NH₂-

aal-aa2-His3 sequence^{2,3} and to be involved in the central nervous system function and cancer growth,⁴ Alzheimer's disease,⁵ cation- π electron interactions in proteins (e.g., Cu²⁺ against tryptophan indole ring⁶), *E. coli* hydrogenases function,^{7,8} targeted cleavage of HIV Rev response element RNA,⁹ calmodulin-peptide complexes.¹⁰ In addition, these motifs are important for the new chemical nuclease design in biotechnology and also as therapeutic agents.^{11,12} The N-terminus region of ATCUN-containing proteins is highly disordered and the geometrical features cannot be easily extracted from the protein structures. The motif participates in the metal interaction with the free N-terminal NH₂ group from residue *aal*, the next two peptide nitrogen atoms from residues *aa2* and His3, and a nitrogen from the imidazole group of His3. In the case of the simulated copper-binding peptide Gly-Gly-His-N-methyl amide, the four nitrogen atoms form a distorted square planar arrangement.¹³ Sankaramakrishnan, Verma, and Kumar¹⁴ reported a list of ATCUN-like motifs from 1949 polypeptide chains and found that only ~1.9 and ~0.3% of histidines are associated with partial and full ATCUN-like geometric features,

* To whom correspondence should be addressed. Phone, (+34) 981 167 000, Ext. 1302; fax, (+34) 981 167 160; e-mail, muntisa@gmail.com; e-mail (H.G.-D.), humberto.gonzalez@usc.es.

[†] University of A Coruña.

[‡] Department of Inorganic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela.

[§] Department of Microbiology & Parasitology, Faculty of Pharmacy, University of Santiago de Compostela.

respectively. They observed that the ATCUN-like motifs are not presented in the middle of the α -helix or β -strand.

The present work uses the protein Quantitative Structure Activity Relationship (QSAR)¹⁵ method for predicting the antitumor activity of ATCUN proteins. We can use many physicochemical parameters such as charges or hydrophilicity parameters,^{16,17} to characterize proteins in these studies. However, many of these QSAR models are based on more simple numerical parameters derived from a graph or network representation of the molecular systems. There are many types of graph representations but essentially they contain two elements: (1) the nodes which are the parts of the system represented by a dot (atoms, amino acids, nucleotides, codons, genes, proteins) and (2) the links between these parts represented as edges or arcs (chemical bonds, hydrogen bonds, reactions, coexpression, regulation and other ties or relationships).^{18–26} Many authors named the numerical parameters used to characterize a graph, which are graph invariants in almost cases, as Topological Indices.^{27–35} This graphic approach of the biological systems study can provide useful insights in QSAR studies,^{36–38} protein functions/attributes^{39–42} or localization,⁴³ protein folding kinetics,⁴⁴ enzyme-catalyzed reactions,^{45–48} inhibition kinetics of processive nucleic acid polymerases and nucleases,^{49–53} DNA sequence analysis,⁵⁴ antisense strand base frequencies,⁵⁵ and analysis of codon usage.^{56,57}

Our research group used the following stochastic molecular descriptors in biochemistry and medicinal chemistry: the entropies,⁵⁸ the spectral moments,⁵⁹ the free energies,^{60,61} and the electrostatic potentials.^{62,63} All these QSAR studies are based on the Markov model (MM) to derive the molecular descriptors that encode the macromolecular structure.⁶⁴ The electrostatic spectral moments were selected for this work by considering the high efficiency shown for protein QSAR models in biochemistry.⁶⁵ We propose the simplest up-to-date reported QSAR equation for the ATCUN antitumor proteins. The average 3D electrostatic spectral moments (π_k) were calculated for 415 proteins, including 133 potential ATCUN antitumor proteins. The Linear Discriminant Analysis model used these TIs to assign proteins into two groups, the ATCUN DNA cleavage proteins (metal-bound active proteins) and the nonactive proteins (metal nonbound inactive proteins). The desirability analysis was used to predict the combined values for the electrostatic spectral moments in the inner region with respect to the total structure that ensures ATCUN mediated anticancer action. In addition, we developed a Receiver Operating Characteristic (ROC) curve analysis to demonstrate that the present model shows significant differences with respect to a random classifier. We demonstrated the robustness of the model by plotting the residuals and explained the model domain applicability by using the model leverage. The results were compared with a similar QSAR model based on the average 3D electrostatic potentials (ξ_k) for ATCUN proteins.³⁶ The ATCUN motifs have been reported to be important for humans,^{66,67} fish⁶⁸ or viruses⁹ but there is no link to parasites. Thus, the Protein Data Bank (PDB) proteins from different parasites were predicted for the DNA cleavage anticancer property by using our best model.

Materials and Methods

Markov Model. The information about the molecular structure of the proteins was codified by using the MM method with the ${}^1\Pi$ matrix (the short-term electrostatic interaction matrix). ${}^1\Pi$ was constructed as a squared matrix ($n \times n$), where n is the number of amino acids (aa) in the protein.^{63,65,69} We considered the hypothetical situation in which every j^{th} - aa has an electrostatic potential φ_j at an arbitrary initial time (t_0). All

the aa_i can interact with electrostatic energy ${}^1E_{ij}$ with every other aa_j in the protein.^{70,71} To simplify the evaluation, a truncation function a_{ij} was applied in such a way that a short-term electrostatic interaction takes place in a first approximation only between neighboring aa ($\alpha_{ij} = 1$). Otherwise, the electrostatic interaction is banished ($\alpha_{ij} = 0$). Thus, the electrostatic interactions propagate indirectly between those aa within the protein backbone, the long-range interactions being possible (not forbidden) and estimated indirectly using the natural powers of ${}^1\Pi$. The spectral moments (π_k) of ${}^1\Pi$ encode information about protein spatial electrostatic indirect interactions between any aa_j and other aa_i one located at a distance k within the 3D protein backbone.^{65,72}

$$\pi_k(O) = \sum_{i=j \in R}^n k p_{ij} = Tr[({}^1\Pi)^k] \quad (1)$$

Equation 1 shows that the present electrostatic spectral moments π_k depend on probabilities $k p_{ij}$ with which the amino acids interact with the other amino acids that are located at a distance $i + [1, 2, 3, \dots k]$. O represents the 3D orbits (regions) of the protein structure where the interacting amino acids are located. By expanding this equation, we can obtain for $k = 0$ the initial unperturbed spectral moments (π_0), for $k = 1$ the short-range (π_1), for $k = 2$ the middle-range (π_2), and for $k = 3$ the long-range spectral moments (π_3), respectively. The notation of the type ' $i + [3,4,5]$ ' refers to the expansion of the descriptors in a series of k indices that encode structural features in the vicinity of the aa_i and is principally used for chain-like data structures such as sequences. This enumeration in the present work refers to sterically close neighbors placed at 1-, 2-, 3-, ... or k -times the 3D cutoff distance. The expansion of eq 1 is illustrated for the tripeptide Ala-Val-Trp (AVW)^{62,63,65,71} in the following equations:

$$\pi_0 = Tr[({}^1\Pi)^0] = Tr\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right) = 3 \quad (2a)$$

$$\pi_1 = Tr[({}^1\Pi)^1] = Tr\left(\begin{bmatrix} {}^1p_{11} & {}^1p_{12} & 0 \\ {}^1p_{21} & {}^1p_{22} & {}^1p_{23} \\ 0 & {}^1p_{32} & {}^1p_{33} \end{bmatrix}\right) = {}^1p_{11} + {}^1p_{22} + {}^1p_{33} \quad (2b)$$

$$\pi_2 = Tr[({}^1\Pi)^2] = Tr\left(\begin{bmatrix} {}^1p_{11} & {}^1p_{12} & 0 \\ {}^1p_{21} & {}^1p_{22} & {}^1p_{23} \\ 0 & {}^1p_{32} & {}^1p_{33} \end{bmatrix} \cdot \begin{bmatrix} {}^1p_{11} & {}^1p_{12} & 0 \\ {}^1p_{21} & {}^1p_{22} & {}^1p_{23} \\ 0 & {}^1p_{32} & {}^1p_{33} \end{bmatrix}\right) = {}^2p_{11} + {}^2p_{22} + {}^2p_{33} \quad (2c)$$

$$\pi_3 = Tr[({}^1\Pi)^3] = Tr\left(\begin{bmatrix} {}^1p_{11} & {}^1p_{12} & 0 \\ {}^1p_{21} & {}^1p_{22} & {}^1p_{23} \\ 0 & {}^1p_{32} & {}^1p_{33} \end{bmatrix} \cdot \begin{bmatrix} {}^1p_{11} & {}^1p_{12} & 0 \\ {}^1p_{21} & {}^1p_{22} & {}^1p_{23} \\ 0 & {}^1p_{32} & {}^1p_{33} \end{bmatrix} \cdot \begin{bmatrix} {}^1p_{11} & {}^1p_{12} & 0 \\ {}^1p_{21} & {}^1p_{22} & {}^1p_{23} \\ 0 & {}^1p_{32} & {}^1p_{33} \end{bmatrix}\right) = {}^3p_{11} + {}^3p_{22} + {}^3p_{33} \quad (2d)$$

To carry out the calculations referred to in eq 1 and detailed in eqs 2a, 2b, 2c and 2d, the elements (${}^1p_{ij}$) of ${}^1\Pi$ were calculated as:^{65,69}

$$\begin{aligned}
 {}^1p_{ij} &= \frac{\alpha_{ij} \cdot E_{ij}}{\sum_{m=1}^{\delta+1} \alpha_{im} \cdot E_{im}} = \frac{\alpha_{ij} \cdot \frac{q_i \cdot q_j}{d_{ij}^2}}{\sum_{m=1}^{\delta+1} \alpha_{im} \cdot \frac{q_i \cdot q_m}{d_{im}^2}} = \frac{\alpha_{ij} \cdot q_i \cdot \frac{q_j}{d_{ij}^2}}{q_i \cdot \sum_{m=1}^{\delta+1} \alpha_{im} \cdot \frac{q_m}{d_{im}^2}} = \\
 &= \frac{\alpha_{ij} \cdot \frac{q_j}{d_{ij}^2}}{\sum_{m=1}^{\delta+1} \alpha_{im} \cdot \frac{q_m}{d_{im}^2}} = \frac{\alpha_{ij} \cdot \varphi_j}{\sum_{m=1}^{\delta+1} \alpha_{im} \cdot \varphi_m} \quad (3)
 \end{aligned}$$

where q_i and q_j are the electronic charges for the i^{th} -aa and the j^{th} -aa, and the neighborhood relationship (truncation function $\alpha_{ij} = 1$) was turned on if these amino acids participate in a peptide hydrogen bond or $d_{ij} < d_{\text{cutoff}} = 5 \text{ \AA}$.^{65,71} The distance d_{ij} is the Euclidean distance between the C_{α} atoms of the two amino acids; d_{oj} is the distance between the amino acid and the charge center of the protein. All distances were obtained from the x , y , and z coordinates of the amino acids from the PDB files.⁷³

The MM was used to calculate average noninteracting [$\tau_0(O)$], short-range [$\tau_1(O)$], middle-range [$\tau_2(O)$], and long-range electrostatic interaction potentials [$\tau_k(O)$, when $k > 2$] for different protein regions called orbits in the 415 proteins. The 3D space of the protein was imaginary spliced into four regions or orbits such as the core (c), the inner (i), the middle (m), and the outer (o). The core orbit is a sphere that contains all the amino acids having the orbit ratio $r < 25\%$ ($r = [d(j)/d(j)_{\text{max}}] \cdot 100$). $d(j)$ is the distance from the C_{α} of the amino acid j to the center of the protein and $d(j)_{\text{max}}$ represents the larger distance for a C_{α} in the protein. The inner orbit is described by $25\% \leq r < 50\%$, the middle orbit by $50\% \leq r < 75\%$ and the outer orbit by $r \geq 75\%$. Thus, five sets or orbits of amino acids (core, inner, middle, surface, total) and six ranges for the electrostatic interactions (0, 1, 2, 3, 4, 5) were considered for the calculation of a total of thirty ($5 \times 6 = 30$) spectral moments^{65,69,71,72} with BIOMARKS tool⁷⁴ to characterize each of the 415 proteins.

Our formalism is the metal-free model. All the analyzed proteins have a metal in the PDB file but we used for our calculations only the protein geometry. Thus, the current QSAR model may predict that a new protein has ATCUN DNA-cleavage activity only if the protein can bind a metal.

Statistical Analysis. The methodology flowchart from Figure 1 gives details about each step of the present work. The 3D electrostatic moments of all the database proteins obtained by using the PDB files and BIOMARKS tool are the base of the next step, the design of a classification model by statistical analysis. Linear discriminant analysis (LDA) has been chosen as the simplest and fastest method. To decide if a protein is classified as having ATCUN activity or not, we added a variable named *ATCUNactiv* (with values of 1 for active or -1 for inactive) and a cross-validation variable (*Sel*). The independent data test is used by splitting the data at random in a training series (*train*, 75%) used for model construction and a prediction one (*val*, 25%) for model validation. The ATCUN activity of these proteins has been known from the literature and has been the result of experiments.¹³

The best QSAR classification model that links the protein structural properties coded in spectral moments with the ATCUN activity is described by the following formula:

$$\text{DNA-cleavage} = c_0 + \sum_{k=1}^n c_k \cdot \pi_k(O) \quad (4)$$

where DNA-cleavage is the continue score value for the *ATCUN non-ATCUN* classification, $\pi_k(o)$ are the 3D spectral moments with k from 1 to n (the initial unperturbed spectral moments for $k = 0$, the short-range spectral moment for $k = 1$, the middle-range spectral moments for $k = 2$, and the long-range spectral moments for $k = 3$), for the amino acid orbits ($c = \text{core}$, $i = \text{inner}$, $m = \text{middle}$, and $o = \text{outer}$); $c_i - c_n$ are the spectral moment coefficients, n is the number for the indices and c_0 is the independent term.

GDA models quality was determined by examining Wilk's λ statistics, leverage threshold to define the model domain (h), the model significance level (p -level), and canonical regression coefficient (R_c). We also inspected the percentage of good classification, cases/variables ratios, and number of variables to be explored to avoid overfitting or chance correlation. The LDA *Forward* stepwise method was used to find the best model. Thus, the training set of proteins were used to create the model and the validation set to verify if the model can accurately predict the ATCUN activity for new proteins (Figure 1). Other methods such as protein modeling and molecular dynamics, may predict the protein geometry and, therefore, some geometry criteria may predict the ATCUN activity of the proteins. The limitations of these methods for the present problem are the following: they are time-consuming and incomplete. Thus, this work presents a better alternative such as a general, fast and accurate method for the evaluation of the ATCUN activity of new proteins by using only the PDB geometry.

Databases. We used a total of 415 proteins to develop the model. The nonactive proteins were randomly selected from the PDB server⁷³ and the list of the potential ATCUN feature antitumor proteins were obtained from the literature.¹³ The PDB database was also used to select 721 proteins from 47 parasite species (only predicted, not used to train or validate the model). The correspondent 1751 protein chains were tested for the DNA cleavage anticancer property by using the best QSAR model resulted.

Results and Discussion

Model for ATCUN Activity. The protein biological activity in organic and inorganic biochemistry can be predicted by using the protein QSAR models combined with simplified truncated electrostatics.⁷⁵⁻⁷⁹ The present work is based on the electrostatic spectral moments $\pi_k(O)$ for a protein QSAR study of interest in bioinorganic chemistry. LDA was used to find the best QSAR model that can classify new proteins into two groups in the absence of prior information: nonactive or potential ATCUN antitumor proteins. The independent data test was used by splitting the data at random in a training series (75%) used for model construction and a prediction one (25%) for model validation (Figure 1). The initial ATCUN activity information (*ATCUNactiv* variable) has been presented in literature¹³ as the result of the experiments. A previous work has reported the applicability of the LDA in QSAR studies.^{36,80-82} The best QSAR LDA model in this study was described by eq 5 and it was obtained with the *Forward stepwise* method from STATISTICA:⁸³

$$\begin{aligned} \text{DNA-cleavage} &= 0.36 \cdot \pi_2(t) + 0.05 \cdot \pi_0(i) - 7.504 \\ N &= 313R_c = 0.77\lambda = 0.40h = 0.058p < 0.001 \end{aligned} \quad (5)$$

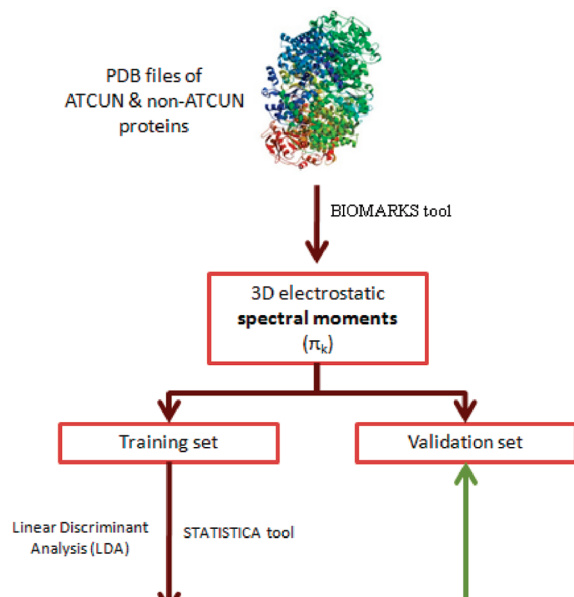


Figure 1. Method flowchart for evaluation of the ATCUN-DNA cleavage activity for new parasite proteins.

where equation elements are $\pi_k(O)$ values with π as the spectral moment, k as the topological distance between the amino acids considered and O between brackets as the orbit of amino acids (i = inner; t = total or whole protein). N represents the number of proteins selected at random from the total amount of 415 and used to train the classification function. The statistical parameters of the same equation were also shown by Wilk's statistic (λ), canonical regression coefficient (R_c), leverage threshold value to define the model domain (h) and the model significance level (p -level).⁸⁴ The model showed excellent accuracy in the training series and predictability in the validation series with an overall good classification of 91.32% (373 out of 415 proteins). The classification matrices for the training, validation and both series are presented in Table 1. The model can be freely used at our Bio-AIMS portal (<http://miaja.tic.udc.es/Bio-AIMS/ATCUNPred.php>).

The proteins can act by diverse mechanisms with different level of effectiveness. For this reason, an ideal QSAR model should be based on quantitative biological activities (e.g., IC_{50}). Even if we do not have these values, we know which proteins present a certain biological activity and which of them do not show any activity. The advantage of using LDA against the regression technique is to be the first method acting as a pattern recognition technique that identifies potentially active proteins and gives a score for the probability of the presence of such activity without predicting how high this probability

Table 1. QSAR Classification Results for Training, Validation, and Both Series

	train			CV			both		
	%	(+)	(-)	%	(+)	(-)	%	(+)	(-)
MM Spectral moments ($n = 415$)									
ATCUN (+)	74.2	72	25	71.1	27	11	73.3	99	36
Nonactive (-)	100.0	0	216	100.0	0	64	100.0	0	280
total	92.0			89.2			91.3		
Electrostatic potential ($n = 265$)									
ATCUN (+)	90.0	90	10	96.9	32	1	91.7	122	11
Nonactive (-)	92.9	7	92	87.8	4	29	91.6	11	121
total	91.5			92.4			91.7		

is.⁸⁴ The misclassified proteins can be explained by the fact that the biological activity of proteins is determined by several forces such as the hydrophobic ones. These proteins are not a representative percentage, only 8.6% of the entire database (36 out of 415).

The coefficients of our best model (eq 5) are standardized and permit comparison and interpretation of the participation of each protein region in the biological activity. Thus, our best model allocates positive contributions of 0.36 to the ATCUN-mediated DNA cleavage activity for unitary increment in the total amount of electrostatic spectral moments $\pi_2(t) = \pi_2(c) + \pi_2(i) + \pi_2(m) + \pi_2(o)$. The catalytic nature of the metal Cu(II)–Ni(II) cluster is explained by the contribution of all the

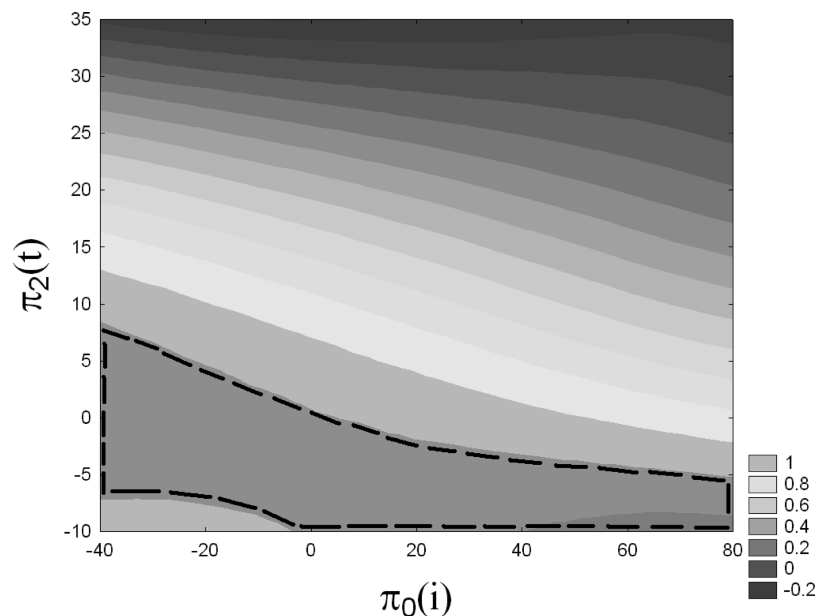


Figure 2. Activity desirability analysis for the classification model variables (orbits or regions).

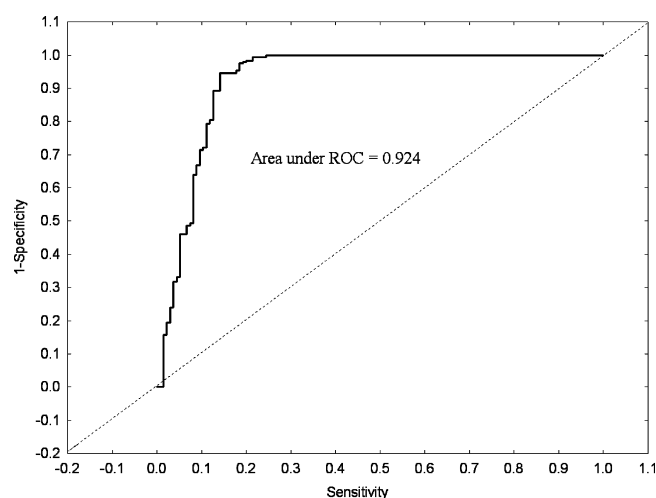


Figure 3. ROC-curve analysis of the DNA cleavage mediated anticancer activity model.

protein orbits even if these motifs were funded at the end of α -helices, at the beginning of the beta-strands and in the turn/random coil regions, but not in the middle of helices/strands. We based this study on the idea that the presence of the ATCUN motif is necessary but not enough for the anticancer activity. Thus, the entire 3D structure of the protein participates in the protein activity because it can influence the accessibility of the ATCUN-like motif, the supra-molecular recognition of the DNA, the subcellular location of the proteins, active site hydrophobicity or other factors. It may also explain the additional positive but lower contribution of 0.05 for the unitary increment in inner spectral moments $\pi_0(i)$. The desirability profiles⁸³ present the levels of the predictor variables $\pi_2(t)$ and $\pi_0(i)$ that produce the most desirable predicted DNA cleavage-mediated anticancer responses (see Figure 2). We can observe that all proteins with $-5 < \pi_2(t) < 5$ (standardized values) are expected to present higher DNA cleavage mediated anticancer activity by accommodation in their backbone of ATCUN motifs for lower values of $\pi_0(i)$ (the region encircled by a white dashed line is wider on the left than on the right).

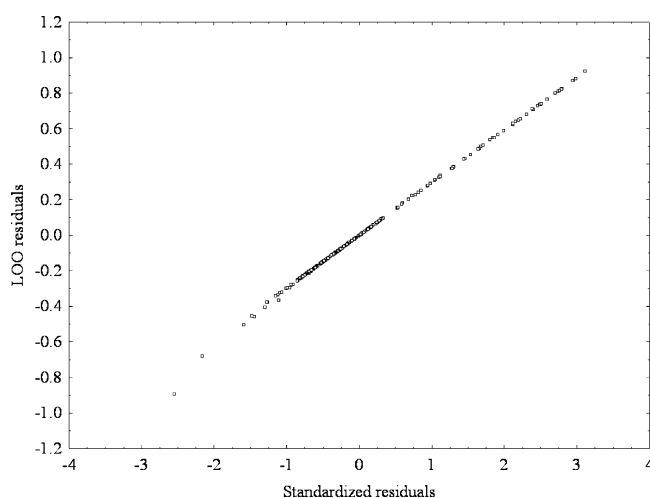


Figure 4. Model robustness to LOO data variation.

The present model (eq 5) is simpler than the previous one (eq 6) reported with the same series of ATCUN proteins.³⁶ The older model was fit using the electrostatic potentials $\xi_k(O)$ of different orbits as described in the following equation:

$$\text{DNA-cleavage} = 1.15 \cdot \xi_1(i) + 2.18 \cdot \xi_5(m) + 27.57 \cdot \xi_0(o) - 27.57 \cdot \xi_0(t) + 0.09N = 199R_c = 0.74\lambda = 0.44p < 0.001 \quad (6)$$

Equation 6 shows higher percentages of good classification for ATCUN proteins but it uses four parameters, which means two times more variables than the model reported in the present work. In addition, the best model with only two $\xi_k(O)$ values classifies worse than the present model with two $\pi_k(O)$. In addition, this model (eq 5) is based on a data set of 313 proteins, which is 1.5 times larger than the one used in the previous model containing 115 proteins (eq 6). Other disadvantage of the previous model is the calculation of the $\xi_k(O)$ values is more complicated whereas the spectral moments $\pi_k(O)$ are straightforward calculated from the traces of matrices.

To check the quality of our model based on complex network spectral moments, we carried out some statistical analysis. The

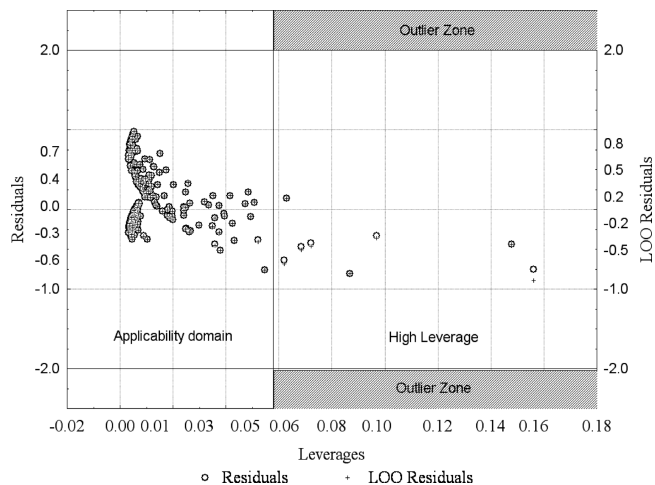


Figure 5. LDA model domain analysis.

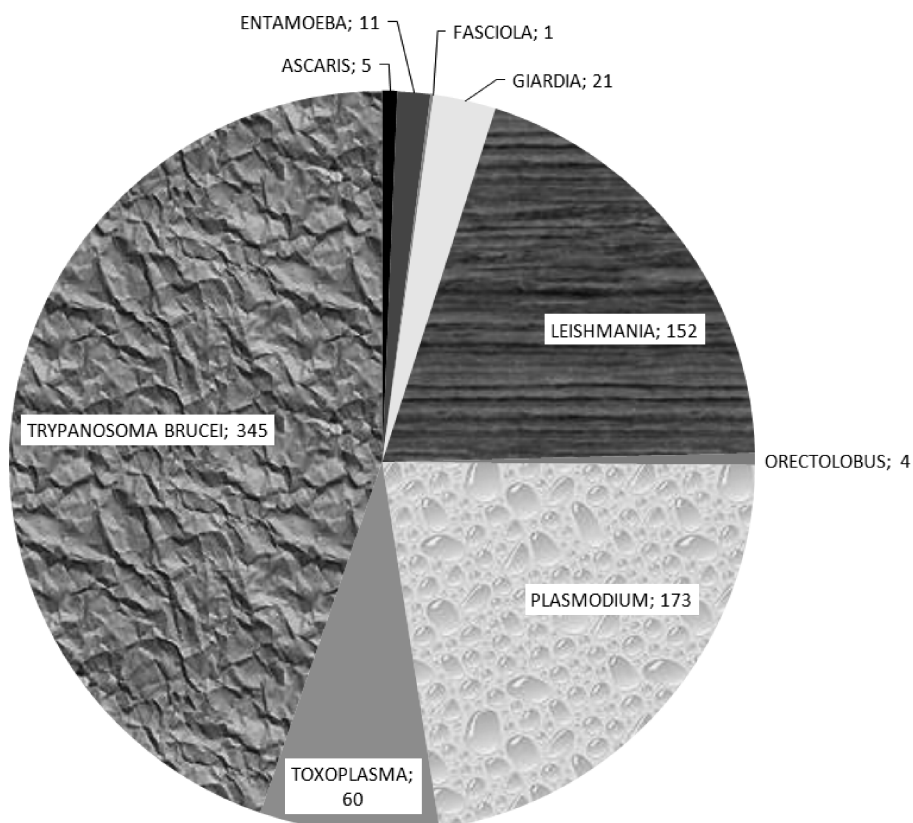


Figure 6. Predicted ATCUN protein chains by parasite family.

ROC-curve analysis tested whether the model behave as a random classifier or not. Random classifiers may be plotted as in a straight line ROC-curve with a 45° slope and an area under the curve equal to 0.5. Conversely, nonrandom classifiers are statistically significant models with an area under the curve above 1. As it can be noted in Figure 3, our model behaves clearly as a not-random statistically significant classifier with an area under the curve of 0.92.⁸⁵

Due to the robustness of the LDA multivariate statistical techniques, the predictive ability and interference reached by using the final model should not be affected (see Figure 4). The linear relationship between the leave-one-out (LOO) residuals and the standardized raw residuals illustrate the high stability of the model to data variation. Finally, we have studied

the Domain of Applicability (DA) of the model due to the natural limitations inherent to QSAR models caused by data conformation. DA may be reduced due to the low number of samples used for training. The simplest method to determine the DA of our QSAR model is the visual inspection of the leverage plot (residuals vs leverages of the training instances).^{86,87} The leverage (h) of a sample in the original variable space measures its influence on the model and it is defined as follows:

$$h_i = \chi_i^T (X^T X)^{-1} \chi_i \quad (i = 1, \dots, n) \quad (7)$$

χ_i are the indices or descriptor vectors of the considered instance (π_k in this work) and X is the model matrix derived

Table 2. DNA Cleavage Evaluation for Parasite Protein Chains

PDB	chain	function	prob. ^a (%)	PDB	chain	function	prob. ^a (%)
<i>Ascaris Suum</i>				<i>Plasmodium falciparum</i>			
1O0S	A	Oxidoreductase	100.000	3EBG	A	Hydrolase	100.000
1LLQ	A	Oxidoreductase	100.000	3EBH	A	Hydrolase	100.000
1LLQ	B	Oxidoreductase	100.000	1ZRO	A	Cell invasion	100.000
1O0S	B	Oxidoreductase	100.000	1ZRO	B	Cell invasion	100.000
1F34	A	Hydrolase	99.172	3EBI	A	Hydrolase	100.000
2BJR	A	Motility	98.982	1ZRL	A	Cell invasion	100.000
2BJR	B	Motility	98.931	2EPH	C	Lyase	100.000
2BJQ	A	Motility	98.640	2PC4	C	Lyase	100.000
1EAI	A	Serine proteinase	95.212	2PC4	A	Lyase	100.000
1EAI	B	Serine proteinase	95.083	2W40	A	Transferase	100.000
<i>Entamoeba Histolytica</i>				<i>Toxoplasma gondii</i>			
2OUI	B	Oxidoreductase	99.9998	3GG8	C	Transferase	100.0000
2OUI	C	Oxidoreductase	99.9997	3GG8	A	Transferase	100.0000
2OUI	D	Oxidoreductase	99.9990	2ABS	A	Signaling	100.0000
2OUI	A	Oxidoreductase	99.9969	3EOE	B	Transferase	100.0000
1Y9A	C	Oxidoreductase	99.9659	3EOE	A	Transferase	100.0000
1Y9A	A	Oxidoreductase	99.9631	3EOE	C	Transferase	100.0000
1OF9	A	Toxin	99.6656	3GG8	B	Transferase	100.0000
1M6J	B	Isomerase	94.7758	3GG8	D	Transferase	100.0000
1M6J	A	Isomerase	94.2867	2JH1	A	Cell adhesion	99.9999
3EMU	A	Hydrolase	88.6560	2AAO	A	Transferase	99.9976
<i>Giardia intestinalis</i>				<i>Trypanosoma brucei</i>			
2FFL	B	Hydrolase	100.0000	1H6Z	A	Transferase	100.0000
2QVW	A	Hydrolase	100.0000	3F5M	C	Transferase	100.0000
2QVW	B	Hydrolase	100.0000	3F5M	A	Transferase	100.0000
2FFL	A	Hydrolase	100.0000	1PGJ	B	Oxidoreductase	100.0000
2FFL	C	Hydrolase	100.0000	1PGJ	A	Oxidoreductase	100.0000
2QVW	C	Hydrolase	100.0000	3F5M	B	Transferase	100.0000
2FFL	D	Hydrolase	100.0000	3F5M	D	Transferase	100.0000
2QVW	D	Hydrolase	100.0000	2HIG	A	Transferase	99.9993
2II2	A	Metal binding	99.9239	2HIG	B	Transferase	99.9989
3GAY	A	Lyase	99.5204	1YAR	U	Hydrolase	99.9985
<i>Leishmania major</i>				<i>Fasciola Hepatica</i>			
2OEF	A	Transferase	100.0000	2O6X	A	Hydrolase	99.9111
2VOB	A	Ligase	100.0000	2VIM	A	Oxidoreductase	85.5820
2VOB	B	Ligase	100.0000	2FHE	B	Transferase	72.2195
2OEG	A	Transferase	100.0000	2FHE	A	Transferase	71.8592
3G1U	B	Hydrolase	100.0000	1FHE	A	Transferase	70.4894
3G1U	D	Hydrolase	99.9999	2FHE	H	Transferase	0.25620
3HJC	A	Chaperone	99.9999	2FHE	G	Transferase	0.25620
3G1U	A	Hydrolase	99.9999				
2VPM	A	Ligase	99.9997				
1OKG	A	Transferase	99.9996				

^aNote: Prob. = probability to have DNA cleavage action based on the best obtained spectral moment QSAR model.

from the training set descriptor values. Thus, the warning leverage h^* is defined by eq 8.

$$h^* = 3 \times p' / n \quad (8)$$

n is the number of training instances and p' is the number of model adjusting parameters. Figure 5 shows the applicability domain of the LDA model, which is determined by training instances with h values lower than $h^* = 0.058$. New samples with an h value higher than h^* and/or a value of standardized residual higher than 2 or lower than -2 are out of the DA bandwidth of the model and, consequently, cannot be reliably predicted.^{88,89}

Predicting ATCUN Proteins in Parasites. The lack of information about the ATCUN motifs in parasites leads to a necessity of testing the parasite proteins with the best resulted

model to evaluate possible DNA cleavage proteins.^{11,12,14,90}

Figure 6 presents the number of the possible ATCUN-like proteins in 9 parasite families with a probability greater than 99%. A large number of protein chains in protozoa such as Trypanosoma, Plasmodium, Leishmania or Toxoplasma have been predicted to present DNA cleavage activity (see Figure 1). The percentages of these highly predicted protein chains from the analyzed ones in all parasites, arranged according to the most important biological function, are the following: 70.5% for oxidoreductases, 62.5% for signaling proteins, 58.2% for lyases, 45.5% for membrane proteins, 44.4% for ligase, 41.3% for hydrolases, 39.2% for transferases, 34.5% for cell adhesion proteins, 33.5% for metal binders, 25.0% for translation proteins, 16.7% for transporters, 9.1% of the structural proteins, and 8.2% for isomerases. From among

these candidates, several chains are pointed out: 2FFL (chains A, B, C, D) as a specialized ribonuclease Dicer that initiates RNA interference by cleaving double-stranded RNA substrates,⁹¹ 2A0U (chains A and B) as a translation initial factor in *Leishmania major*, 2II2 (chain A), 3CHJ (chain A), 3CHL (chain A) as a member of the alpha giardin family of annexins localized to the flagella of the intestinal protozoan parasite *Giardia lamblia*^{92,93} and 3CS1 (chain A) as the flagellar calcium-binding protein (FCaBP) of the protozoan *Trypanosoma cruzi*.⁹⁴ In addition, a protein with unknown biological function is predicted to have DNA cleavage activity (1N81, 186 amino acids, *Plasmodium falciparum*). For more detailed information, Table 2 presents the top ten of the best predicted ATCUN proteins in eight important parasites. We can observe different protein functions of the predicted protein chains such as oxidoreductase for *Ascaris suum* and *Entamoeba histolytica*, transferase for *Toxoplasma gondii*, *Trypanosoma brucei* and *Fasciola hepatica*, hydrolase for *Giardia intestinalis* and *Leishmania major*, and lyase for *Plasmodium falciparum*.

In general, most methods that predict protein functions are reliant on identifying a similar protein and transferring its annotations to the query protein. An example is the BLAST⁹⁵ method that fails when a similar protein cannot be identified, or when any similar proteins identified also lack reliable annotations.⁹⁶ At the moment, there is no template of ATCUN protein in the BLAST server and therefore the BLAST method fails to predict the ATCUN DNA-cleavage activity of proteins. As an advantage, the current method can predict the ATCUN function of a protein even if it has other known activity.

Conclusions

The study of the metal–protein functions and interactions is a topic of great importance, and several authors have presented interesting results.^{97,98} The present work proposes a new QSAR model based on the electrostatic spectral moment indices and evaluates the presence of the potential ATCUN-like antitumor activity of the proteins. All of the calculations have been made using the 3D structure information contained in PDB files for metal–unbound or free proteins, and the resulting model is simpler compared with a similar model based on the electrostatic potentials.³⁶ Thus, the present QSAR approach is very useful in bioinorganic chemistry for the prediction of the biological activity of potential metal–protein complexes whose free protein structure has been characterized but the metal interactions remain unexplored. The desirability analysis of the model predicts the values for the spectral moments in one single region for the ATCUN-like proteins. The evaluation of the DNA cleavage activity for the parasite protein chains by using the present web implemented model was preceded and became a starting point for future experimental and theoretical studies of parasite pathologies.

Acknowledgment. C.R.M. and H.G.-D., from the Faculty of Computer Science, University of A Coruña and the Faculty of Pharmacy, University of Santiago de Compostela (Spain), respectively, acknowledge financial support granted by Isidro Parga Pondal program of Xunta de Galicia. We also thank the General Directorate of Scientific and Technologic Promotion of the Galician University System of the Xunta de Galicia for the grants 2007/127 and 2007/144.

References

- Laussac, J. P.; Sarkar, B. Characterization of the copper(II)- and nickel(II)-transport site of human serum albumin. Studies of copper(II) and nickel(II) binding to peptide 1–24 of human serum albumin by ¹³C and ¹H NMR spectroscopy. *Biochemistry* **1984**, *23* (12), 2832–8.
- Kimoto, E.; Tanaka, H.; Gytoku, J.; Morishige, F.; Pauling, L. Enhancement of antitumor activity of ascorbate against Ehrlich ascites tumor cells by the copper:glycylglycylhistidine complex. *Cancer Res.* **1983**, *43* (2), 824–8.
- Jin, Y.; Lewis, M. A.; Gokhale, N. H.; Long, E. C.; Cowan, J. A. Influence of stereochemistry and redox potentials on the single- and double-strand DNA cleavage efficiency of Cu(II) and Ni(II) Lys-Gly-His-derived ATCUN metallopeptides. *J. Am. Chem. Soc.* **2007**, *129* (26), 8353–61.
- Harford, C.; Sarkar, B. Neuromedin C binds Cu(II) and Ni(II) via the ATCUN motif: implications for the CNS and cancer growth. *Biochem. Biophys. Res. Commun.* **1995**, *209* (3), 877–82.
- Drew, S. C.; Noble, C. J.; Masters, C. L.; Hanson, G. R.; Barnham, K. J. Pleomorphic copper coordination by Alzheimer's disease amyloid-beta peptide. *J. Am. Chem. Soc.* **2009**, *131* (3), 1195–207.
- Yorita, H.; Otomo, K.; Hiramatsu, H.; Toyama, A.; Miura, T.; Takeuchi, H. Evidence for the cation- π interaction between Cu²⁺ and tryptophan. *J. Am. Chem. Soc.* **2008**, *130* (46), 15266–7.
- Dias, A. V.; Mulvihill, C. M.; Leach, M. R.; Pickering, I. J.; George, G. N.; Zamble, D. B. Structural and biological analysis of the metal sites of *Escherichia coli* hydrogenase accessory protein HypB. *Biochemistry* **2008**, *47* (46), 11981–91.
- Chung, K. C.; Cao, L.; Dias, A. V.; Pickering, I. J.; George, G. N.; Zamble, D. B. A high-affinity metal-binding peptide from *Escherichia coli* HypB. *J. Am. Chem. Soc.* **2008**, *130* (43), 14056–7.
- Jin, Y.; Cowan, J. A. Targeted cleavage of HIV rev response element RNA by metallopeptide complexes. *J. Am. Chem. Soc.* **2006**, *128* (2), 410–1.
- Mal, T. K.; Ikura, M.; Kay, L. E. The ATCUN domain as a probe of intermolecular interactions: application to calmodulin-peptide complexes. *J. Am. Chem. Soc.* **2002**, *124* (47), 14002–3.
- Singh, R. K.; Sharma, N. K.; Prasad, R.; Singh, U. P. DNA cleavage study using copper (II)-GlyAibHis: a tripeptide complex based on ATCUN peptide motifs. *Protein Pept. Lett.* **2008**, *15* (1), 13–9.
- Melino, S.; Gallo, M.; Trotta, E.; Mondello, F.; Paci, M.; Petruzzelli, R. Metal-binding and nuclease activity of an antimicrobial peptide analogue of the salivary histatin 5. *Biochemistry* **2006**, *45* (51), 15373–83.
- Harford, C.; Sarkar, B. Amino Terminal Cu(II)- and Ni(II)-Binding (ATCUN) Motif of Proteins and Peptides: Metal Binding, DNA Cleavage, and Other Properties. *Acc. Chem. Res.* **1997**, *30* (3), 123–30.
- Sankaramakrishnan, R.; Verma, S.; Kumar, S. ATCUN-like metal-binding motifs in proteins: identification and characterization by crystal structure and sequence analysis. *Proteins* **2005**, *58* (1), 211–21.
- Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach: The Netherlands, 1999.
- Zbilut, J. P.; Giuliani, A.; Colosimo, A.; Mitchell, J. C.; Colafranceschi, M.; Marwan, N.; Webber, C. L., Jr.; Uversky, V. N. Charge and hydrophobicity patterning along the sequence predicts the folding mechanism and aggregation of proteins: a computational approach. *J. Proteome Res.* **2004**, *3* (6), 1243–53.
- Shen, B.; Bai, J.; V, M. Physicochemical feature-based classification of amino acid mutations. *Protein Eng., Des. Sel.* **2008**, *21* (1), 37–44.
- Krishnan, A.; Zbilut, J. P.; Tomita, M.; Giuliani, A. Proteins as networks: usefulness of graph theory in protein science. *Curr. Protein Pept. Sci.* **2008**, *9* (1), 28–38.
- Krishnan, A.; Giuliani, A.; Zbilut, J. P.; Tomita, M. Implications from a network-based topological analysis of ubiquitin unfolding simulations. *PLoS ONE* **2008**, *3* (5), e2149.
- Palumbo, M. C.; Colosimo, A.; Giuliani, A.; Farina, L. Essentiality is an emergent property of metabolic network wiring. *FEBS Lett.* **2007**, *581* (13), 2485–9.
- Krishnan, A.; Giuliani, A.; Tomita, M. Indeterminacy of reverse engineering of Gene Regulatory Networks: the curse of gene elasticity. *PLoS ONE* **2007**, *2* (6), e562.
- Tun, K.; Dhar, P. K.; Palumbo, M. C.; Giuliani, A. Metabolic pathways variability and sequence/networks comparisons. *BMC Bioinformatics* **2006**, *7*, 24.
- Nandy, A.; Ghosh, A.; Nandy, P. Numerical characterization of protein sequences and application to voltage-gated sodium channel α subunit phylogeny. *In Silico Biol.* **2009**, *9*, 8.

- (24) Randić, M.; Vracko, M.; Nandy, A.; Basak, S. C. On 3-D graphical representation of DNA primary sequences and their numerical characterization. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (5), 1235–44.
- (25) Nandy, A.; Basak, S. C. Simple numerical descriptor for quantifying effect of toxic substances on DNA sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (4), 915–9.
- (26) Nandy, A.; Basak, S. C.; Gute, B. D. Graphical representation and numerical characterization of H5N1 avian flu neuraminidase gene sequence. *J. Chem. Inf. Model* **2007**, *47* (3), 945–51.
- (27) Liao, B.; Wang, T. M. Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1666–70.
- (28) Liao, B.; Ding, K. Graphical approach to analyzing DNA sequences. *J. Comput. Chem.* **2005**, *26* (14), 1519–23.
- (29) Randić, M. Condensed representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (1), 50–6.
- (30) Randić, M.; Vraško, M.; Nandy, A.; Basak, S. C. On 3-D Graphical Representation of DNA Primary Sequences and Their Numerical Characterization. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235–44.
- (31) Randić, M.; Basak, S. C. Characterization of DNA primary sequences based on the average distances between bases. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 561–8.
- (32) Randić, M.; Balaban, A. T. On a four-dimensional representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 532–9.
- (33) Bielińska-Waż, D.; Nowak, W.; Waz, P.; Nandy, A.; Clark, T. Distribution Moments of 2D-graphs as Descriptors of DNA Sequences. *Chem. Phys. Lett.* **2007**, *443*, 408–13.
- (34) Agüero-Chapin, G.; Gonzalez-Diaz, H.; Molina, R.; Varona-Santos, J.; Uriarte, E.; Gonzalez-Diaz, Y. Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from *Psidium guajava* L. *FEBS Lett.* **2006**, *580*, 723–30.
- (35) Krishnan, A.; Giuliani, A.; Zbilut, J. P.; Tomita, M. Network scaling invariants help to elucidate basic topological principles of proteins. *J. Proteome Res.* **2007**, *6* (10), 3924–34.
- (36) González-Díaz, H.; Sanchez-Gonzalez, A.; Gonzalez-Diaz, Y. 3D-QSAR study for DNA cleavage proteins with a potential anti-tumor ATCUN-like motif. *J. Inorg. Biochem.* **2006**, *100* (7), 1290–7.
- (37) González-Díaz, H.; Bonet, I.; Terán, C.; de Clercq, E.; Bello, R.; García, M.; Santana, L.; Uriarte, E. ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. *Eur. J. Med. Chem.* **2007**, *42*, 580–5.
- (38) Prado-Prado, F. J.; González-Díaz, H.; Martínez de la Vega, O.; 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*, 5871–80.
- (39) Munteanu, C. R.; Gonzalez-Diaz, H.; Magalhaes, A. L. Enzymes/non-enzymes classification model complexity based on composition, sequence, 3D and topological indices. *J. Theor. Biol.* **2008**, *254* (2), 476–82.
- (40) Munteanu, C. R.; Gonzalez-Diaz, H.; Borges, F.; de Magalhaes, A. L. Natural/random protein classification models based on star network topological indices. *J. Theor. Biol.* **2008**, *254* (4), 775–83.
- (41) Xiao, X.; Chou, K. C. Digital coding of amino acids based on hydrophobic index. *Protein Pept. Lett.* **2007**, *14* (9), 871–5.
- (42) Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Chou, K. C. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino Acids* **2006**, *30* (1), 49–54.
- (43) Nair, R.; Rost, B. LOC3D: annotate sub-cellular localization for protein structures. *Nucleic Acids Res.* **2003**, *31* (13), 3337–40.
- (44) Chou, K. C. Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. *Biophys. Chem.* **1990**, *35*, 1–24.
- (45) Chou, K. C. Graphical rules in steady and non-steady enzyme kinetics. *J. Biol. Chem.* **1989**, *264*, 12074–12079.
- (46) Chou, K. C.; Forsen, S. Graphical rules for enzyme-catalyzed rate laws. *Biochem. J.* **1980**, *187*, 829–835.
- (47) Chou, K. C.; Liu, W. M. Graphical rules for non-steady state enzyme kinetics. *J. Theor. Biol.* **1981**, *91* (4), 637–54.
- (48) Kuzmic, P.; Ng, K. Y.; Heath, T. D. Mixtures of tight-binding enzyme inhibitors. Kinetic analysis by a recursive rate equation. *Anal. Biochem.* **1992**, *200*, 68–73.
- (49) Althaus, I. W.; Chou, J. J.; Gonzales, A. J.; Diebel, M. R.; Chou, K. C.; Kezdy, F. J.; Romero, D. L.; Aristoff, P. A.; Tarpley, W. G.; Reusser, F. Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. *J. Biol. Chem.* **1993**, *268*, 6119–6124.
- (50) Althaus, I. W.; Chou, J. J.; Gonzales, A. J.; Diebel, M. R.; Chou, K. C.; Kezdy, F. J.; Romero, D. L.; Aristoff, P. A.; Tarpley, W. G.; Reusser, F. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. *Biochemistry* **1993**, *32*, 6548–6554.
- (51) Althaus, I. W.; Chou, J. J.; Gonzales, A. J.; LeMay, R. J.; Deibel, M. R.; Chou, K. C.; Kezdy, F. J.; Romero, D. L.; Thomas, R. C.; Aristoff, P. A. Steady-state kinetic studies with the polysulfonate U-9843, an HIV reverse transcriptase inhibitor. *Experientia* **1994**, *50* (1), 23–8.
- (52) Althaus, I. W.; Chou, K. C.; Lemay, R. J.; Franks, K. M.; Deibel, M. R.; Kezdy, F. J.; Resnick, L.; Busso, M. E.; So, A. G.; Downey, K. M.; Romero, D. L.; Thomas, R. C.; Aristoff, P. A.; Tarpley, W. G.; Reusser, F. The benzylthio-pyrimidine U-31,355, a potent inhibitor of HIV-1 reverse transcriptase. *Biochem. Pharmacol.* **1996**, *51* (6), 743–50.
- (53) Chou, K. C.; Kezdy, F. J.; Reusser, F. Review: Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. *Anal. Biochem.* **1994**, *221*, 217–230.
- (54) Qi, X. Q.; Wen, J.; Qi, Z. H. New 3D graphical representation of DNA sequence based on dual nucleotides. *J. Theor. Biol.* **2007**, *249* (4), 681–90.
- (55) Chou, K. C.; Zhang, C. T.; Elrod, D. W. Do “antisense proteins” exist. *J. Protein Chem.* **1996**, *15* (1), 59–61.
- (56) Chou, K. C.; Zhang, C. T. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. *AIDS Res. Hum. Retroviruses* **1992**, *8*, 1967–76.
- (57) Zhang, C. T.; Chou, K. C.; Analysis of codon usage in 1562, *E. Coli* protein coding sequences. *J. Mol. Biol.* **1994**, *238*, 1–8.
- (58) Ramos de Armas, R.; González-Díaz, H.; Molina, R.; Uriarte, E. Markovian Backbone Negentropies: Molecular descriptors for protein research. I. Predicting protein stability in Arc repressor mutants. *Proteins* **2004**, *56* (4), 715–23.
- (59) González-Díaz, H.; Uriarte, E. Biopolymer stochastic moments. I. Modeling human rhinovirus cellular recognition with protein surface electrostatic moments. *Biopolymers* **2005**, *77* (5), 296–303.
- (60) González-Díaz, H.; Agüero, G.; Cabrera, M. A.; Molina, R.; Santana, L.; Uriarte, E.; Delogu, G.; Castanedo, N. Unified Markov thermodynamics based on stochastic forms to classify drugs considering molecular structure, partition system, and biological species: distribution of the antimicrobial G1 on rat tissues. *Bioorg. Med. Chem. Lett.* **2005**, *15* (3), 551–7.
- (61) González-Díaz, H.; Cruz-Monteagudo, M.; Molina, R.; Tenorio, E.; Uriarte, E. Predicting multiple drugs side effects with a general drug-target interaction thermodynamic Markov model. *Bioorg. Med. Chem.* **2005**, *13* (4), 1119–29.
- (62) Gonzalez-Diaz, H.; Molina, R.; Uriarte, E. Recognition of stable protein mutants with 3D stochastic average electrostatic potentials. *FEBS Lett.* **2005**, *579* (20), 4297–301.
- (63) González-Díaz, H.; Pérez-Bello, A.; Uriarte, E. Stochastic molecular descriptors for polymers. 3. Markov electrostatic moments as polymer 2D-folding descriptors: RNA-QSAR for mycobacterial promoters. *Polymer* **2005**, *46*, 6461–73.
- (64) Freund, J. A.; Poschel, T. Stochastic Processes in Physics, Chemistry, and Biology. In *Lecture Notes in Physics*; Springer-Verlag: Berlin, Germany, 2000.
- (65) González-Díaz, H.; Uriarte, E.; Ramos de Armas, R. Predicting stability of Arc repressor mutants with protein stochastic moments. *Bioorg. Med. Chem.* **2005**, *13* (2), 323–31.
- (66) Gasmí, G.; Singer, A.; Forman-Kay, J.; Sarkar, B. NMR structure of neuromedin C, a neurotransmitter with an amino terminal CuII-NiII-binding (ATCUN) motif. *J. Pept. Res.* **1997**, *49* (6), 500–9.
- (67) Gokhale, N. H.; Cowan, J. A. Inactivation of human angiotensin converting enzyme by copper peptide complexes containing ATCUN motifs. *Chem Commun (Camb)* **2005**, (47), 5916–8.
- (68) Robertson, L. S.; Iwanowicz, L. R.; Marranca, J. M. Identification of centrarchid hepcidins and evidence that 17 β -estradiol disrupts constitutive expression of hepcidin-1 and inducible expression of hepcidin-2 in largemouth bass (*Micropterus salmoides*). *Fish Shellfish Immunol* **2009**, *26* (6), 898–907.
- (69) Saiz-Urra, L.; González-Díaz, H.; Uriarte, E. Proteins Markovian 3D-QSAR with spherically-truncated average electrostatic potentials. *Bioorg. Med. Chem.* **2005**, *13* (11), 3641–7.
- (70) González-Díaz, H.; Molina, R.; Uriarte, E. Stochastic molecular descriptors for polymers. 1. Modelling the properties of icosahedral viruses with 3D-Markovian negentropies. *Polymer* **2003**, (45), 3845–53.
- (71) González-Díaz, H.; Molina, R.; Uriarte, E. Markov entropy backbone electrostatic descriptors for predicting proteins biological activity. *Bioorg. Med. Chem. Lett.* **2004**, *14* (18), 4691–5.
- (72) González-Díaz, H.; Saiz-Urra, L.; Molina, R.; Uriarte, E. Stochastic molecular descriptors for polymers. 2. Spherical truncation of

- electrostatic interactions on entropy based polymers 3D-QSAR. *Polymer* **2005**, *46*, 2791–8.
- (73) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–42.
- (74) González-Díaz, H.; Molina, R., BIOMARKS version 1.0, contact information: gonzalezdiaz@yahoo.es or qohumbe@usc.es2005.
- (75) Kundu, S.; Gupta-Bhaya, P. How a repulsive charge distribution becomes attractive and stabilized by a polarizable protein dielectric. *J. Mol. Struct. (Theochem)* **2004**, 668.
- (76) Burykin, A.; Warshel, A. On the origin of the electrostatic barrier for proton transport in aquaporin. *FEBS Lett.* **2004**, *570* (1–3), 41–6.
- (77) Norberg, J.; Nilsson, L. On the truncation of long-range electrostatic interactions in DNA. *Biophys. J.* **2000**, *79* (3), 1537–53.
- (78) Navarro, E.; Fenude, E.; Celda, B. Conformational and structural analysis of the equilibrium between single- and double-strand beta-helix of a D, L-alternating oligonucleotide. *Biopolymers* **2004**, *73* (2), 229–41.
- (79) Costa, L. A.; Rocha, W. R.; De Almeida, W. B.; Dos Santos, H. F. Linear free energy relationship for 4-substituted (o-phenylenediamine)platinum(II) dichloride derivatives using quantum mechanical descriptors. *J. Inorg. Biochem.* **2005**, *99* (2), 575–83.
- (80) Perez Gonzalez, M.; Morales Helguera, A. TOPS-MODE versus DRAGON descriptors to predict permeability coefficients through low-density polyethylene. *J. Comput. Aided Mol. Des.* **2003**, *17* (10), 665–72.
- (81) Marrero-Ponce, Y.; Medina-Marrero, R.; Torrens, F.; Martinez, Y.; Romero-Zaldivar, V.; Castro, E. A. Atom, atom-type, and total nonstochastic and stochastic quadratic fingerprints: a promising approach for modeling of antibacterial activity. *Bioorg. Med. Chem.* **2005**, *13* (8), 2881–99.
- (82) Marrero-Ponce, Y.; Montero-Torres, A.; Zaldivar, C. R.; Veitia, M. I.; Perez, M. M.; Sanchez, R. N. Non-stochastic and stochastic linear indices of the 'molecular pseudograph's atom adjacency matrix': application to 'in silico' studies for the rational discovery of new antimalarial compounds. *Bioorg. Med. Chem.* **2005**, *13* (4), 1293–304.
- (83) STATISTICA, (data analysis software system), version 6.0, www.statsoft.com; StatSoft, Inc., 2002.
- (84) Van Waterbeemd, H., Chemometric methods in molecular design. In *Method and Principles in Medicinal Chemistry*; Manhnhold, R., Krogsgaard-Larsen, P., Timmerman, H., Van Waterbeemd, H., Eds.; Wiley-VCH: New York, 1995; Vol. 2, pp 283–93.
- (85) González-Díaz, H.; Vina, D.; Santana, L.; de Clercq, E.; Uriarte, E. Stochastic entropy QSAR for the in silico discovery of anticancer compounds: prediction, synthesis, and in vitro assay of new purine carbanucleosides. *Bioorg. Med. Chem.* **2006**, *14* (4), 1095–107.
- (86) Atkinson, A. C., *Plots, Transformations, and regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*; Clarendon Press: Oxford, 1985.
- (87) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T.; McDowell, R. M.; Gramatica, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* **2003**, *111* (10), 1361–75.
- (88) Monari, G.; Dreyfus, G. Local overfitting control via leverages. *Neural Comput.* **2002**, *14* (6), 1481–506.
- (89) Meloun, M.; Syrový, T.; Bordovská, S.; Vrana, A. Reliability and uncertainty in the estimation of pKa by least squares nonlinear regression analysis of multiwavelength spectrophotometric pH titration data. *Anal. Bioanal. Chem.* **2007**, *387* (3), 941–55.
- (90) Melino, S.; Garlando, L.; Patamia, M.; Paci, M.; Petruzzelli, R. A metal-binding site is present in the amino terminal region of the bioactive iron regulator hepcidin-25. *J. Pept. Res.* **2005**, *66* (s1), 65–71.
- (91) Macrae, I. J.; Zhou, K.; Li, F.; Repic, A.; Brooks, A. N.; Cande, W. Z.; Adams, P. D.; Doudna, J. A. Structural basis for double-stranded RNA processing by Dicer. *Science* **2006**, *311* (5758), 195–8.
- (92) Pathuri, P.; Nguyen, E. T.; Svard, S. G.; Luecke, H. Apo and calcium-bound crystal structures of Alpha-11 giardin, an unusual annexin from *Giardia lamblia*. *J. Mol. Biol.* **2007**, *368* (2), 493–508.
- (93) Pathuri, P.; Nguyen, E. T.; Ozorowski, G.; Svard, S. G.; Luecke, H. Apo and calcium-bound crystal structures of cytoskeletal protein alpha-14 giardin (annexin E1) from the intestinal protozoan parasite *Giardia lamblia*. *J. Mol. Biol.* **2009**, *385* (4), 1098–112.
- (94) Wingard, J. N.; Ladner, J.; Vanarotti, M.; Fisher, A. J.; Robinson, H.; Buchanan, K. T.; Engman, D. M.; Ames, J. B. Structural insights into membrane targeting by the flagellar calcium-binding protein (FCaBP), a myristoylated and palmitoylated calcium sensor in *Trypanosoma cruzi*. *J. Biol. Chem.* **2008**, *283* (34), 23388–96.
- (95) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215* (3), 403–10.
- (96) Dobson, P. D.; Doig, A. J. Predicting enzyme class from protein structure without alignments. *J. Mol. Biol.* **2005**, *345* (1), 187–99.
- (97) Di Cera, E. Thrombin: a paradigm for enzymes allosterically activated by monovalent cations. *C R Biol.* **2004**, *327* (12), 1065–76.
- (98) Nayal, M.; Di Cera, E. Valence screening of water in protein crystals reveals potential Na⁺ binding sites. *J. Mol. Biol.* **1996**, *256* (2), 228–34.

PR900556G

LIBP-Pred: web server for lipid binding proteins using structural network parameters; PDB mining of human cancer biomarkers and drug targets in parasites and bacteria†Humberto González-Díaz,^{*a} Cristian R. Munteanu,^b Lucian Postelnicu,^c Francisco Prado-Prado,^d Marcos Gestal^b and Alejandro Pazos^b

Received 19th October 2011, Accepted 1st December 2011

DOI: 10.1039/c2mb05432a

Lipid-Binding Proteins (LIBPs) or Fatty Acid-Binding Proteins (FABPs) play an important role in many diseases such as different types of cancer, kidney injury, atherosclerosis, diabetes, intestinal ischemia and parasitic infections. Thus, the computational methods that can predict LIBPs based on 3D structure parameters became a goal of major importance for drug-target discovery, vaccine design and biomarker selection. In addition, the Protein Data Bank (PDB) contains 3000+ protein 3D structures with unknown function. This list, as well as new experimental outcomes in proteomics research, is a very interesting source to discover relevant proteins, including LIBPs. However, to the best of our knowledge, there are no general models to predict new LIBPs based on 3D structures. We developed new Quantitative Structure–Activity Relationship (QSAR) models based on 3D electrostatic parameters of 1801 different proteins, including 801 LIBPs. We calculated these electrostatic parameters with the MARCH-INSIDE software and they correspond to the entire protein or to specific protein regions named core, inner, middle, and surface. We used these parameters as inputs to develop a simple Linear Discriminant Analysis (LDA) classifier to discriminate 3D structure of LIBPs from other proteins. We implemented this predictor in the web server named LIBP-Pred, freely available at <http://miaja.tic.udc.es/Bio-AIMS/LIBPpred.php>, along with other important web servers of the Bio-AIMS portal. The users can carry out an automatic retrieval of protein structures from PDB or upload their custom protein structural models from their disk created with LOMETS server. We demonstrated the PDB mining option performing a predictive study of 2000+ proteins with unknown function. Interesting results regarding the discovery of new Cancer Biomarkers in humans or drug targets in parasites have been discussed here in this sense.

Introduction

Fatty Acid-Binding Proteins (FABPs) or, generally speaking, Lipid-Binding proteins (LIBPs) play important roles in many diseases. The mammalian FABPs bind long-chain FA with high affinity. The recent discussion carried out by Storch and McDermott¹ highlights that the large number of FABP types is suggestive of distinct functions in specific tissues. Thus, the

LIBPs modulate intracellular lipid homeostasis by regulating FA transport in the nuclear and extra-nuclear compartments of the cell; in doing so, they also impact systemic energy homeostasis. In this sense, the characterization of LIBPs has become important for vaccine design, drug-target discovery, and disease biomarkers selection. Noiri and Doi *et al.*² have reported that urinary FABP 1 as an early predictive biomarker of kidney injury and a liver-type LIBP are included in a panel of biomarkers in acute and chronic kidney disease.³ Evennett and Petrov *et al.*⁴ discussed that the performance of the currently available serological markers is suboptimal for routine clinical use, but novel markers of intestinal ischemia such as i-FABP may offer improved diagnostic accuracy. Krusinova and Pelikanova⁵ reviewed adipocyte/macrophage FABP (A-FABP) that has been shown to be closely associated with metabolic syndrome, obesity and development of atherosclerosis and has been recently suggested as a potential therapeutic target of these abnormalities in animal models. New agents in development for the treatment of bacterial infections include LIBPs inhibitors.⁶

^a Department of Microbiology & Parasitology, Faculty of Pharmacy, University of Santiago de Compostela, Praza Seminario de Estudos Galegos, s/n. Campus Sur, 15782 Santiago de Compostela, Spain. E-mail: gonzalezdiazh@yahoo.es

^b Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruña, 15071 A Coruña, Spain

^c S.C. POLIPHARMA INDUSTRIES S.R.L., 550052 Sibiu, Romania

^d Department of Organic Chemistry, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c2mb05432a

LIBPs are also very relevant for different types of cancer. Liver FABP (L-FABP) is a new prognostic factor for hepatic resection of colorectal cancer metastases, FABP-6⁷ is also overexpressed in colorectal cancer, and the overexpression of FABP-7⁸ correlates with basal-like subtype of breast cancer. There have been studies on the fatty acid metabolism in human breast cancer cells (MCF7) transfected with heart-type FABP (H-FABP)⁹ and the role of sterol regulatory element binding protein-1c¹⁰ in the regulation of fatty acid synthase expression in breast cancer. Li and Huang *et al.*¹¹ carried out a prognostic evaluation of epidermal FABP and calyphosine, two proteins implicated in endometrial cancer using a proteomic approach. Cutaneous FABP (C-FABP)¹² expressed in prostate cancer is a potential prognostic marker and target for tumorigenicity-suppression and Adipocyte-FABP (A-FABP)¹³ induces apoptosis in DU145 prostate cancer cells. In addition, Hammamieh *et al.*¹⁴ evaluated *in vitro* molecular impacts of antisense complementary to the FABP mRNA in DU145 prostate cancer cells.

On the other hand, LIBPs or FABPs are also very important in parasites. Three different classes of small LBPs are found in helminth parasites. The parasites that produce these proteins are unable to synthesize their own complex lipids and, instead, rely entirely upon their hosts for supply.¹⁵ Zhu¹⁶ has reviewed fatty acid metabolism in *Cryptosporidium parvum*, which is one of the apicomplexans that can cause severe diarrhea in humans and animals. The slow development of anti-cryptosporidiosis chemotherapy is primarily due to the poor understanding of the basic metabolic pathways in this parasite. Many well-defined or promising drug targets found in other apicomplexans are either absent or highly divergent in *C. parvum*. The recently discovered apicoplast and its associated Type II fatty acid synthetic enzymes in *Plasmodium* sp., *Toxoplasma* sp., and *Eimeria* sp. apicomplexans are absent in *C. parvum*, suggesting that this parasite is unable to synthesize fatty acids *de novo*. However, *C. parvum* possesses other important LIBPs enzymes involved in fatty acid metabolism.¹⁷ In addition, molecular cloning of components of protective antigenic preparations has suggested that related parasite LIBPs could form the basis of the protective immune cross-reactivity between the parasitic trematode worms *Fasciola hepatica* and *Schistosoma mansoni*. Tendler and Brito¹⁸ discussed that these results suggest a single vaccine effective against at least two parasites, *F. hepatica* and *S. mansoni*, of veterinary and human importance, respectively. In fact, schistosomes are the causative agents of schistosomiasis, one of the most prevalent and serious parasitic diseases, that currently affects approximately 200 million people worldwide. Schistosome excretory/secretory (ES) proteins have been shown to play important roles in modulating mammalian host immune systems. In parallel, Liu *et al.*¹⁹ performed a global proteomics identification of the ES proteins from adult worms of *Schistosoma japonicum*, one of the three major schistosome species. They revealed that LIBPs are major constituents of the *in vitro* ES proteome. Actually, in the 1990s, WHO/TDR created a product development programme and initiated collaborations with other major international donors to promote rapid vaccine development and other tools for the control of endemic diseases. The LIBP-14 kDa antigen of *S. mansoni* (Sm14) stands out, both due to its steady progress towards field trials

and because it represents the sole vaccine candidate to emerge from an endemic country. Studies have now progressed to the scale-up level and an industrial production process has successfully been put in place. It has been demonstrated that it is effective not only against *S. mansoni* in humans, but also against *F. hepatica*, a parasite that causes disease in cattle and sheep leading to annual losses over 3 \$US billion to the food industry worldwide. The Sm14 patents have been granted to Oswaldo Cruz Foundation (FIOCRUZ),¹⁸ a Brazilian scientific institution directly linked to the Brazilian Ministry of Health. In fact, free-living nematodes, such as *Caenorhabditis elegans*, also secrete a structurally novel class of proteins (FARs) that present both FAB and retinol-binding activity into the surrounding tissues of the host. One important class of FARs is the nematode polyprotein allergens/antigens (NPAs),²⁰ these proteins are of interest because they may play an important role in scavenging fatty acids and retinoids from the host that are essential for the survival of the parasite and also because the localised depletion of such lipids may have immunomodulatory effects that compromise the host immune response.

Since fatty acids are essential components of all bio-membranes, molecular and functional studies on LIBPs point new directions for the drug-target discovery, vaccine design, or biomarker prediction for many human metabolic and other diseases, as well as against parasitic diseases. In any case, the number of proteins of different organisms to be experimentally assayed is so vast that the use of computational techniques may be of help to speed-up the process. For instance, very recently Kuang and Colgrave *et al.*²¹ have revealed the complexity of the secreted NPA and FAR FABPs families of *Haemonchus contortus* by an iterative proteomics–bioinformatics approach. The parasite *H. contortus*, also known as red stomach worm, wire worm or Barber's pole worm, is a very common parasite and one of the most pathogenic nematodes of ruminants. Using the human genome database, the recently developed G-protein-coupled receptor (GPCR) deorphanization strategy has successfully identified multiple LIBPs receptors for fatty acids.²²

On the other hand, we can use, in principle, structure-dependent physicochemical parameters, such as charges or hydrophilicity parameters,^{23,24} to characterize proteins in quantitative structure–function relationship studies, also known as Quantitative Structure–Activity Relationships (QSAR).²⁵ However, many of these QSAR models are based on more simple numerical parameters, called Topological Indices (TIs),²⁶ derived from a graph or network representation of the molecular systems (including but not limited to protein structure, as in this case). In fact, there are many types of graph representations, but essentially they contain two elements: (1) the nodes which are the parts of the system represented by a dot (atoms, amino acids, nucleotides, codons, genes, proteins, metabolites, ... *etc.*) and (2) the links between these parts represented as edges or arcs (chemical bonds, hydrogen bonds, metabolic reactions, co-expression, regulation and other ties or relationships).^{27–37} In any case, with the generalization of Internet, the development of new predictive methods has become the first step in the application of computational techniques to proteome research. Nowadays, it is not sufficient to develop a fast and accurate

predictive model, we should also implement it in public servers, preferably of free access, for the use of the scientific community.³⁸ The server packages developed by Chou and Shen to predict the function of proteins from structural parameters or explore protein structures^{39–42} are good examples in this sense. These may be used by proteome research scientists through interacting with user-friendly interfaces. It means that the user does not need to be an expert on the theoretical details behind this kind of models, including the vast literature published by Chou *et al.* on the development of models with pseudo-amino acid composition parameters or the use of machine learning classification techniques and other algorithms.^{43–48} In any case, to the best of our knowledge, in the literature there is no other theoretical method to predict LIBPs in parasites, cancer tissue, or other disease-specific proteomes that are not present in humans or other organisms, based on the 3D structure of proteins.

According to a recent comprehensive review,⁴⁹ to establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web-server for the predictor that is accessible to the public. Let us describe as follows how to deal with these steps one by one.

González-Díaz *et al.* introduced the method called MARKovian CHemicals *IN Silico* DEsign (MARCH-INSIDE 1.0) for the computational design of small-sized drugs. The approach uses a Markov Chain model (MCM) of the intra-molecular movement of electrons to calculate structural parameters of drugs. In successive studies, we have extended this method to perform fast calculation of 2D and 3D alignment-free structural parameters based on molecular vibrations in RNA secondary structures, or electrostatic potential, and van der Waals interactions in proteins. Recently, the method has been renamed as MARKov CHains Invariants for Networks SIMulation and DEsign (MARCH-INSIDE 2.0). This explores more adequately the broad uses of the method that describes the structure of drugs,⁵⁰ RNA,⁵¹ and proteins,^{52–54} as well as drug–drug networks,⁵⁵ drug–protein interactions.⁵⁶ The MARCH-INSIDE may also be used to study PPIs, bacteria–bacteria co-aggregation, parasite–host interactions and other systems with an MCM associated to a network. In very recent reviews, we have discussed the last applications of this method.^{57–61} We should also make reference to the recent implementation carried out by Munteanu and González-Díaz of the Internet portal called Bio-AIMS, freely available for the use of the international research community. This portal includes the web-server packages TargetPred (<http://bio-aims.udc.es/TargetPred.php>) with new Protein-QSAR servers based on MARCH-INSIDE. One of the servers is ATCUNPred,⁶² useful for predicting ATCUN-mediated DNA-cleavage anti-cancer proteins. The second server is EnzClassPred,⁶³ which implements one of the MARCH-INSIDE-based QSAR models for the prediction of enzyme function.⁶⁴ Two additional servers based on MARCH-INSIDE are: Trypano-PPI⁶⁵ and Plasmod-PPI.⁶⁶ These are the first servers that predict self-protein–protein

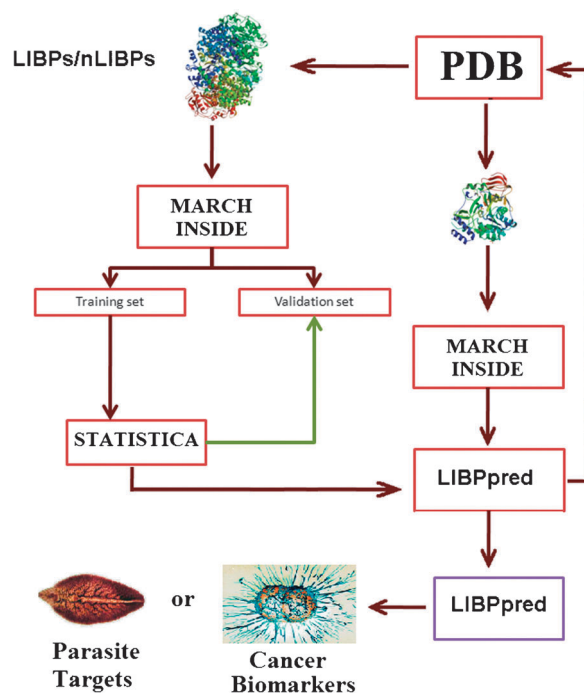


Fig. 1 Flowchart for all the steps necessary to construct/use the classifiers and server.

complexes in *Trypanosome* sp. or *Plasmodium* sp. proteomes, opening new opportunities for anti-trypanosome or anti-malarial drug target discovery.

For all these reasons, we use the MARCH-INSIDE approach in this work to solve the problem of predicting LIBPs from the 3D structure of proteins. In the present work, we have developed the first 3D-QSAR method useful to discriminate between LIBP and non-LIBPs (nLIBPs). Using MARCH-INSIDE 2.0 we have calculated different local and global parameters to a large series of LIBPs and nLIBPs (see Fig. 1). The parameters calculated are of three different classes: average electrostatic potentials $\xi_k(\mathbf{R})$, together with spectral moments of $\pi_k(\mathbf{R})$ and entropy measures $\theta_k(\mathbf{R})$ of the electrostatic field of amino acids placed at distance k from each other within different regions \mathbf{R} of the protein 3D structure. Next, we have carried out a statistical analysis in order to seek a linear equation (3D-QSAR model) that links the 3D electrostatic parameters of the protein structural network with S(LIBP) values. The S(LIBP) output is a real-valued variable that scores the propensity of a protein to act as a LIBP. In addition, we have implemented the model in a public web server for the prediction of these proteins called LIBP-Pred. Last, we have illustrated the use of LIBP-Pred to carry out online data mining of the PDB. We have predicted S(LIBP) values for 2000+ proteins in humans and parasites with known structure but unknown function. This type of study may help us to discover new LIBPs useful as human cancer biomarkers of drug targets in parasites.

Materials and methods

Computational methods

MARCH-INSIDE method. In this work, the information about the molecular structure of the proteins is codified by

using the MM method with the ${}^1\Pi$ matrix (the short-term electrostatic interaction matrix). The matrix ${}^1\Pi$ is constructed as a squared matrix ($n \times n$), where n is the number of amino acids (aa) in the protein.^{67–69} In previous works we have predicted protein function based on $\mu_k(\mathbf{R})$ and $\theta_k(\mathbf{R})$ values of 3D-potentials for different types of interactions or molecular fields derived from ${}^1\Pi$. The main types of the used molecular fields are: E, vdW, and HINT potentials.^{53,68,70} In this paper, we have calculated $\pi_k(\mathbf{R})$ and $\theta_k(\mathbf{R})$ values only for E and HINT potentials. We have omitted the vdW term due to a simple reason; the HINT potential includes a vdW component. The values have been used here as inputs to construct the QSAR model. The detailed explanation has been published before. As follows, we give the formula for $\pi_k(\mathbf{R})$, $\theta_k(\mathbf{R})$ and $\xi_k(\mathbf{R})$ and some general explanations:

$$\xi_k(\mathbf{R}) = - \sum_{j \in \mathbf{R}}^k p_j(\mathbf{R}) \xi_0(j) \quad (1)$$

$$\theta_k(\mathbf{R}) = - \sum_{j \in \mathbf{R}}^n p_j(\mathbf{R}) \log[p_j(\mathbf{R})] \quad (2)$$

$$\pi_k(\mathbf{R}) = \sum_{i=j \in \mathbf{R}}^n p_{ij}(\mathbf{R}) \quad (3)$$

It is remarkable that the spectral moments depend on the probability ${}^k p_{ij}(\mathbf{R})$ with which the effect of the interaction f propagates from amino acid i th to other neighbouring amino acids j th and returns to i th after k -steps. On the other hand, both the average electrostatic potential and the entropy measures depend on the absolute probabilities ${}^k p_f(\mathbf{R})$ with which the amino acid j th has an interaction of type f with the rest of the amino acids. In any case, both probabilities refer to a first ($k = 1$) direct interaction of type f between amino acids placed at a distance equal to k -times the cut-off distance ($r_{ij} = kr_{\text{cut-off}}$). The method uses a Markov Chain Model (MCM) to calculate these probabilities; which also depend on the 3D interactions between all pairs of amino acids placed at a distance r_{ij} in 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 neighbouring 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 (see Fig. 2). In this network the nodes are the C_α atoms of the amino acids and the edges connect pairs of amino acids with $\alpha_{ij} = 1$. Euclidean 3D space $r_3 = (x, y, z)$ coordinates of the C_α atoms of amino acids are listed in protein PDB files. For calculation all water molecules and metal ions were removed.⁵⁸ All calculations were carried out with our in-house MARCH-INSIDE 2.0 software.⁵⁸ For calculation the MARCH-INSIDE software never 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 or orbitals (\mathbf{R}). These regions are often defined in geometric terms and called core, inner, middle or surface region. The protein is virtually divided into the following regions: c corresponds to core, i to inner, m to middle, and s to surface regions, respectively. The diameters of the regions,

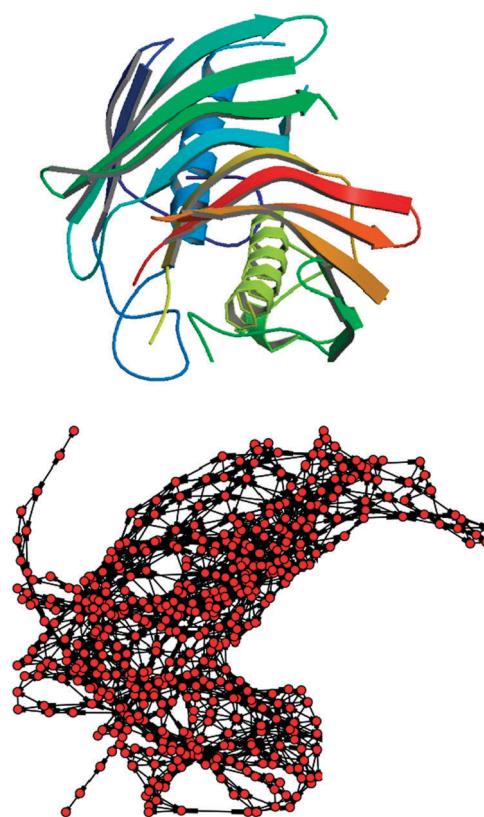


Fig. 2 Representations of a LIBP with PDB ID 1ZHG (an FABP from *P. falciparum*): (A) 3D structure model for full complex and (B) complex network graph for chain A.

as a percentage of the longest distance r_{max} with respect to the centre of charge, are 0 to 25 for region c , 25 to 50 for region i , 50 to 75 for region m , and 75 to 100 for region s . Additionally, we consider the total region (t) that contains all the amino acids in the protein (region diameter 0 to 100% of r_{max}). Consequently, we can calculate different $\pi_k(f)$ and $\theta_k(f)$ for the amino acids contained in a region (c, i, m, s , or t) and placed at a topological distance k within this region (k is the name of the order).^{53,71–74} In this work, we calculated a total of 90 indices (3 types of indices \times 5 types of regions \times 6 higher order considered) for each protein.

LDA model. Linear Discriminant Analysis (LDA) is frequently used for classification/prediction problems in physical anthropology, but it is unusual to find examples in which researchers consider the statistical limitations and assumptions required for this technique. In this work, all LDA models have been trained with the STATISTICA 6.0 software, for which our laboratory holds rights of use.⁷⁵ In LDA, we use several variable selection techniques to seek the model: (i) *all effects* (include all parameters), (ii) *forward-stepwise*, (iii) *forward-entry*, (iv) *backward-stepwise*, (v) *backward-removal*, and (vi) *best subsets*. Unless we specify a different value, we always set a prior probability of $p(\text{LIBP}) = p(\text{nLIBP}) = 0.5$. The LDA discriminant equation was obtained using as input the three types of Markov chain invariants $\theta_k(\mathbf{R})$. The general form of the equation obtained by LDA is:

$$S(\text{LIBP}) = \sum_{\mathbf{R}, k, i}^{5,5,3} a_{\mathbf{R},k} \xi_k(\mathbf{R}) + b_{\mathbf{R},k} \theta_k(\mathbf{R}) + c_{\mathbf{R},k} \pi_k(\mathbf{R}) + d \quad (4)$$

S(LIBP) is the above-mentioned output of the model. It is a real-valued variable that scores the propensity of a protein to act as a LIBP. The χ^2 and p -level values were examined in order to test the statistical significance of the model. The Accuracy, Specificity, Sensitivity were used to quantify the goodness-of-fit and the discriminatory power of the model. Different authors have applied this type of LDA model using different classes of input variables to construct QSAR models for drugs,^{76–86} proteins or nucleic acids.^{80–82,87–91}

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.⁹² However, out of the three test methods, the jackknife test is deemed the most objective.⁹³ 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.⁹² (ii) For the subsampling test, the concrete procedure usually used in the literature 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 demonstrated by eqn (28)–(30) in ref. 49. 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 the same benchmark dataset and the 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 the 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 with strong math background to examine the quality of various predictors (see, e.g. ref. 94–103). However, to reduce the computational time, in this study we have adopted the independent testing dataset cross-validation as many investigators had done with SVM as the prediction engine.

Dataset. The protein structures were downloaded from PDB¹⁰⁴ using the following schemes for PDB-database search: (i) introducing as input parameter the text “fatty acid-binding” in the search item called function for positive cases. Scheme (ii)

was used to get negative cases introducing the PDB IDs for all the proteins contained in the list reported in the article of Dobson and Doig.¹⁰⁵ The positive cases are those proteins with function annotation as LIBPs in the PDB. The list of negative cases of nLIBPs from the search scheme (ii) contains enzymes and other proteins present in humans and many other organisms including other parasites (see ESI 1†). The nLIBPs have known functions different from LIBPs. The dataset was made up of 1801 proteins (801 LIBPs and 1000 nLIBPs) from more than 20 organisms, including parasites and human or cattle hosts. Detailed information about the PDB ID, the values of the electrostatic potential indices, the corresponding observed classification, and the predicted classification for each protein is given in the ESI 2.†

To avoid homology bias and remove the redundant sequences from the benchmark dataset, a cutoff threshold of 25% was recommended⁴⁹ to exclude those proteins from the benchmark datasets that have equal to or greater than 25% sequence identity compared to any other as done in ref. 94 and 106–108. However, in this study we have not used such a stringent criterion because the currently available data do not allow us to do so. Otherwise, the number of proteins for some subsets would be too low to have statistical significance.

Results and discussion

Alignment-free LDA model for LIBPs

Multiple experimental approaches have shown that individual LIBPs possess both unique and overlapping functions, some of which are based on specific elements in the protein structure. Although FA binding affinities for all LIBPs tend to correlate directly with FA hydrophobicity, structure–function studies indicate that subtle three-dimensional (3D) changes that occur upon ligand binding may promote specific protein–protein or protein–membrane interactions that ultimately determine the function of each LIBP. The conformational changes are focused on the LIBP helical/portal domain, a region that was identified by *in vitro* studies to be vital for the FA transport properties of the LIBPs.¹ In this sense, structural parameters that numerically describe both the global and local 3D structure of proteins may be useful for the study of LIBPs. Previous work has reported the applicability of the LDA in QSAR studies.^{109–112} The best QSAR LDA model in this study is described by eqn (5) and was obtained with the *Forward stepwise* method from STATISTICA:¹¹³

$$\begin{aligned} S(\text{LIBP}) = & 12.851\pi_1(c) + 18.355\pi_4(c) - 27.331\pi_5(c) \\ & + 6.870\pi_3(t) - 5.761\pi_4(i) + 1.510\pi_1(s) - 1.074\pi_2(t) \\ & + 0.292\pi_3(t) + 2.030\pi_4(t) - 5.4601; \\ N = & 1351, R_c = 0.78, \chi^2 = 1259.574, p < 0.001 \quad (5) \end{aligned}$$

Interestingly, only the spectral moments of the electrostatic field are linearly correlated to LIBP/nLIBP discrimination. As mentioned in the Materials and methods section, we have explored three types of input variables to seek this equation: $\zeta_k(\mathbf{R})$, $\pi_k(\mathbf{R})$, and $\theta_k(\mathbf{R})$ values. π indicates spectral moments of the electrostatic field, ζ average electrostatic potentials, and θ entropy values of the electrostatic field. This indicates that

self-return propagation of electrostatic interactions within a protein 3D backbone control LIBP action instead of the magnitude of the electrostatic potential *per se* (ξ potential control) or the total information about electrostatic interactions (θ entropy control). On the other hand, we should note that the model determines different effects (in sign and intensity) over PABP action of different amino acids placed at different distances within different regions of the protein backbone. Remember that parameter k accounts for the topological distance between the amino acids considered and R refers to the protein region. Then, we can conclude that according to our model fatty acid-binding seems to be modulated by region-specific propagation of electrostatic interactions within the protein. This effect should be correlated to the physico-chemical mechanism of LIBP action. However, the explanation of this mechanism is a goal beyond the scope of this work, which is oriented to the development of a LIBP predictor and not to unravel the mechanism of action of LIBPs. Consequently, we have focused more on the statistical quality of the model. The statistical parameters of the model are: Canonical Regression Coefficient (R_c), Chi-square (χ^2) and model significance level (p -level).¹¹⁴ N represents only the number of proteins used to train the model. We split the dataset at random in a training series (75%), used for model construction; and a prediction one (25%) used for model validation. The high R_c above 0.8 indicates a strong linear correlation between input and output. The value of p -level < 0.05 for the Chi-square test indicates a statistically significant discrimination between the two groups of proteins. In addition, the model has shown good Accuracy, Specificity, and Sensitivity values in both training series and external validation series. The classification matrices for the training, validation and both series are presented in Table 1. The PDIB, π_k , and S(LIBP) values for all proteins used to train or validate (cv) the model are given in the ESI 2† (available online or upon author's request). This result confirms a statistically significant relationship between MARCH-INSIDE parameters and LIBPs activity. Taking into consideration that this classifier is a simpler linear equation with only nine input parameters we can conclude that this may become a very useful model.

LIBP-Pred web-server

In the Internet era, training and validation of a QSAR and/or computational model should be considered the first step towards the development of a valuable tool for bioinformatics application in proteome research. At the present time, seeking a fast and accurate predictive model is not enough, it should

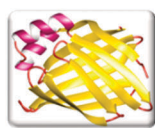
also be implemented into public servers, preferably of free access, available online to the scientific community. The server packages developed by Chou and Shen that predict the function of proteins from structural parameters or explore protein structures^{39–42} are good examples in this sense. These may be used by proteome research scientists by interacting with user-friendly interfaces. It means that the user does not need to be an expert on the theoretical details behind this kind of models, including the vast literature published by Chou *et al.* on the development of models with pseudo-amino acid composition parameters or the use of ML classification techniques and other algorithms.^{43–47} However, to the best of our knowledge, there is no QSAR-based server for the prediction of LIBPs. In this sense, we have implemented the best LDA model found here at the web portal Bio-AIMS as an online server called LIBP-Pred. The acronym LIBP-Pred comes from Lipid Binding Proteins Predictor. LIBP-Pred is located at <http://miaja.tic.udc.es/Bio-AIMS/LIBP-Pred.php>. This online tool is based on PHP/HTML and Python routines coupled to nested MARCH-INSIDE classic algorithm to calculate input molecular structure parameters.⁵⁷

LIBP-Pred mode 1. In Fig. 3, we depict the user interface for LIBP-Pred including mode 1 (top of the web page). The user only has to paste the PDB ID of the query proteins with unknown functions. With these PDB ID codes, LIBP-Pred automatically connects to the PDB database, uploads the PDB files with the 3D structure of the protein, constructs the Markov matrix of electrostatic interactions and calculates the total and region (R) average electrostatic potential values $\pi_k(R)$ for each query protein.

LIBP-Pred mode 2. In mode 1, LIBP-Pred may be used to select potential LIBPs between proteins with known 3D structures that have been released from PDB but with unknown function. However, there are other potential uses of this server. How should one predict S(LIBP) values for proteins with known sequence but unknown 3D structure and function that have not been released to PDB? Mode 2 is essentially the same as mode 1, but the server prompts the users to upload ent and pdb files with 3D structures of proteins generated by using LOMETS web server¹¹⁵ developed by Prof. Zhang *et al.* at Michigan University. In Fig. 3 we depict the user interface for LIBP-Pred mode 2 (bottom of the web page). LOMETS is a local threading meta-server, for quick and automated predictions of protein tertiary structures and spatial constraints. Nine state-of-the-art threading programs are installed and run in a local computer cluster, which ensure the quick generation of initial threading alignments compared to traditional remote-server-based meta-servers. Consensus models are generated from the top predictions of the component-threading servers, which are at least 7% more accurate than the best individual servers based on a TM-score at a t -test significance level of 0.1%. Moreover, side-chain and C-alpha contacts of 42 and 61% accuracy, respectively, as well as long- and short-range distant maps, are automatically constructed from the threading alignments. These data can be easily used as constraints to guide the *ab initio* procedures such as TASSER for further protein tertiary structure modeling. The LOMETS server is

Table 1 Results of the 3D-QSAR study of LIBPs with LDA

Data Sub-set	Group	Parameter	%	nLIBPs	LIBPs
Training	nLIBPs	Specificity	90.0	675	75
	LIBPs	Sensitivity	87.4	76	525
	Total	Accuracy	88.8		
Validation	nLIBPs	Specificity	91.6	229	21
	LIBPs	Sensitivity	88.0	24	176
	Total	Accuracy	90.0		
Both training + validation	nLIBPs	Specificity	90.4	904	96
	LIBPs	Sensitivity	87.5	100	701
	Total	Accuracy	89.1		



LIBPpred

Lipid-Binding Proteins Prediction

Tool: MARCH-INSIDE (Python version)

LDA classification model

Accuracy = 89.11%

(the model is based on 9 spectral moments of the proteins and the final form will be available after the publication)

Note: The LIBP prediction is calculated using $(LIBPscore - Min_score) * 100 / (Max_score - Min_score)$, where LIBPscore is the result of the LDA equation for the current protein and Min and Max score are the minimum and maximum values of the LIBPscore for our dataset.

Mode 1: Standard PDBs

PDB/PDB chain List : Please paste the ID of the PDBs/PDB chains as a list (maximum 10 items)

```
1QGHK
1I4M
2QZTB
1B0U
```

Predict

Data: RCSB PDB

Mode 2: LOMETS PDB

Upload & evaluate one PDB from LOMETS (max. 2MB)

Please select LOMETS PDB

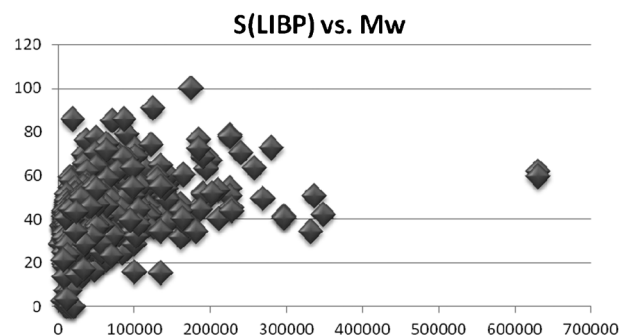
Browse... Predict

Fig. 3 Web-user interface of LIBP-Pred tool.

freely available to the academic community at <http://zhang.bioinformatics.ku.edu/LOMETS>. After generating PDB files with LOMETS we can upload them to LIBP-Pred. This is the same strategy used to develop the mode 2 of the web server MIND-BEST to predict drug–target interactions between drugs and proteins with unknown 3D structure.¹¹⁶ Anyhow, we have to be aware that by using this input mode 2 we can predict S(LIBP) values using 3D structural models generated only by modelling. Consequently, predictions derived with input mode 2 have to be used with higher caution than predictions obtained with input mode 1.

LIBP-Pred mining of PDB

The existence in PDB of 3000+ proteins with unknown function and the interest in the discovery of new LIBPs or LBPs as drug targets in parasite infections or cancer biomarkers prompt us to carry out a data mining search of new LIBPs candidates in PDB. For this study we have implemented the key function PDB mining in the new server LIBP-Pred. By clicking this key the server performs automatic search of all PDB files with unknown function at a reference date. After that, LIBP-Pred extracts all C_{α} coordinates from these files and calculates the necessary $\pi_k(R)$ values for all these proteins. Last, the server uses these values as inputs of the best model found and predicts the S(LIBP) values for all these proteins. The proteins with highest scores may be selected as candidates for experimental assays in order to confirm LIBP function. Each time we use the PDB mining key, the server updates the prediction for all new PDB files present in the last version of the PDB synchronized with LIBPpred. We have predicted S(LBP) values for a total of 2693 proteins selected to have unknown function (or only hypothetical function predicted) and low sequence homology in current PDB release. A total of 552 out of 2693 proteins studied (20.5%) were predicted as possible LIBPs with S(LIBP) > 50%. However, if we restrict the criteria to S(LIBP) > 55% in order to discard unclassified outputs the results shrink to 271 possible LIBPs (10.1%).¹¹⁷ These are in any case “weak”

Fig. 4 Scatter plot of S(LIBP)% vs. molecular weight (M_w) of the protein complex.

criteria somehow; if we use a more restrictive criterion for this LDA classifier with a cut-off of 75% our LIBP-Pred found only 27 possible LIBPs (1%). Another important result is the demonstration that LIBP-Pred predictions are not molecular-weight dependent (biased), see Fig. 4. This scatter plot shows that there are no apparent linear relationships between S(LIBP) and M_w with a correlation coefficient of only $R^2 = 0.079$ between both properties. Consequently, we can conclude that LIBP-Pred takes into consideration specific 3D structural features and not a simply burden M_w -biased predictor.

The value of S(LBP) = 93.87% was the highest value predicted by LIBP-Pred for a protein with unknown function selected out of the 2693 mentioned before. This value corresponds to the chain A of the protein with PDB ID = 2RJB. The protein, deposited in PDB with unknown function, is expressed by *Shigella flexneri*, a bacterium that causes severe dysentery in human beings. This result is very interesting because of the importance of the lipid (*i.e.* phosphoinositide) metabolic pathway in the regulation of cellular processes implicated in survival, motility, and trafficking; which is often subverted by bacterial pathogens. In fact, *S. flexneri* infection has been demonstrated recently to generate the lipid PI5P to alter endocytosis and prevent termination of EGFR signaling.¹¹⁸

This property is used by *S. flexneri* to favour survival of host cells in the infection process. In this sense, if it is finally confirmed as a LIBP, the present results may point out the chain A of 2RJB as a possible target for anti-bacterial drugs effective against this human pathogen.

Mining of parasite proteins in PDB with LIBP-Pred

LIBPs (including FABPs) are being studied as important actors in host–parasite interactions that may become important targets to halt infections caused by pathogen parasites in human beings. For instance, the malaria parasite liver stage produces tens of thousands of red cell-infectious forms within its host hepatocyte. It is thought that the vacuole-enclosed parasite completely depends on the host cell for successful development but the molecular parasite–host cell interactions underlying this remarkable growth have remained elusive. Using a yeast two-hybrid screen and a yeast over-expression system, some authors have shown that UIS3, a parasite protein essential for liver stage development, interacts directly

with liver-fatty acid binding protein, L-FABP. Down-regulation of L-FABP expression in hepatocytes severely impairs parasite growth and over-expression of L-FABP promotes growth. This is the first identified direct liver stage-host cell protein interaction, providing a possible explanation for the importance of UIS3 in liver infection.¹¹⁹ With these facts in mind, we decided to explore the proteins present in *Plasmodium* sp. proteome reported in PDB with known 3D structure but unknown function in order to possibly discover new LIBPs relevant to Malaria disease.

Considering that LIBPs, as well as other LBP, are not exclusive for *Plasmodium* but are also present in other parasites we have used LIBP-Pred to study proteins of other parasites also present in PDB but without function annotation. The highest S(LIBP) values predicted for all proteins studied with unknown function that are expressed in parasites correspond to one protein of *C. parvus* (see Table 2). The PDB IDs and score for this protein are: PDB ID = 2O1OA (2O1O chain A) and S(LBP) = 85.63%. This is a very high value, according to our web server, that may support a more serious inspection of this protein as probable LIBP. 2O1O is a complex protein

Table 2 Top hits of LIBPs predicted in *H. sapiens*, Parasites and other organisms

Species/Organism	PDB ID	S%	M_w	PDB ID	S%	M_w
Top LIBP-Pred hits for different organisms				<i>Leishmania major</i>		
<i>Shigella flexneri</i>	2RJB	93.87	211122.31	1X9GA ^a	51.23	22701.7
<i>Thermus thermophilus</i>	1WDTA	92.03	73780.59	3HA4A	51.12	141446.42
<i>Arthrobacter aurescens</i>	3IUKA	91.01	124991.81	1TC5A	50.84	87100
<i>Neisseria meningitidis</i>	1VGYA	88.96	86192	3M3IA	49.68	202852
<i>Thermus thermophilus</i>	1WDIA	88.73	38675.12	1Y1XA	48.91	43851.29
<i>Shewanella oneidensis</i>	1ZEEA	87.28	93492	3HA4B	48.2	141446.42
<i>Haemophilus influenzae</i>	3M73A	87.2	36346.29	3S4OA	47.6	37372.82
<i>Arabidopsis thaliana</i>	1YDUA	85.72	18979.9	1YF9A	46.45	58695.01
<i>Cryptosporidium parvum</i>	2O1OA	85.63	86659.06	3K5VA	44.96	16307.28
<i>Chlamydomonas reinhardtii</i>	3CE2A	85.01	71228.68	2AR1A	44.59	20443.59
<i>Methanocaldococcus jannaschii</i>	2AEUA	79.24	42903.78	1Y63A	43.68	22570.21
<i>Staphylococcus aureus</i>	1QYIA	78.99	44007.6	3LJNA	40.59	41122.7
<i>Oleispira antarctica</i>	3IRUA	78.96	61467.58	1YQFA	40.5	138256.19
<i>Aquifex aeolicus</i>	2HEKB	78.95	90765.37	1R75A	35.79	16322.39
<i>Plasmodium</i> sp.				<i>Homo sapiens</i>		
<i>vivax</i>	2GUUA	65.9	40489.07	2WM3A	62.66	34893.97
<i>berghei</i>	2FDSA	65.71	82299.41	2GTRA	61.04	87455.7
<i>falciparum</i>	2QU8A	58.77	25964.6	2EC4A	55.37	20134.9
<i>falciparum</i>	1Z40A	55.17	76870.25	2HV6B	53.56	73663.31
<i>falciparum</i>	1Z40E	54.63	76870.25	2HV6A	53.39	73663.31
<i>vivax</i>	2B30A	52.98	137232.12	2FBMA	53.3	96749.27
<i>falciparum</i>	1XQ9A	51.33	59869.23	2Q4KA	53.13	82801.8
<i>knowlesi</i>	1TXJA	49.94	19876.3	2I6TA	52.59	66672.09
<i>falciparum</i>	1Y6ZA	49.79	61130	2O95A	51.35	44000.2
<i>falciparum</i>	1N81A	49.78	22520.9	2P2LA	50.99	64220.55
<i>falciparum</i>	2FBNA	47.09	46076.4	2DB9A	50.83	16665.1
<i>falciparum</i>	3NI8A	45.99	18591.19	1NZNA	50.75	14955.85
<i>vivax</i>	2FO3A	45.19	14333.7	1X53A	50.31	16177.2
<i>falciparum</i>	1TQXA	45.08	51476.48	2L2OA	50.31	10125.7
<i>falciparum</i>	2P65A	44.78	20439.7	2O95B	50.11	44000.2
<i>falciparum</i> 3D7	2H2YA	40.78	62364.8	2P5XA	49.85	51829.34
<i>falciparum</i>	2VWAA	38.07	77780.69	2K07A	48.97	20557.7
<i>falciparum</i>	1SYRA	36.72	153103.19	2DLXA	48.48	17337.5
<i>falciparum</i>	2FU0A	34.59	18188.8	1V9VA	48.01	12426.2
<i>falciparum</i>	2KDNA	33.88	12328.2	1WRYA	47.96	13124.5
Other parasites				<i>Caenorhabditis elegans</i>		
<i>Toxoplasma gondii</i>	2F4ZB	40.76	43101.09	1XKQA	74.13	122794.91
<i>Trypanosoma brucei</i>	2Q0XA	55.26	74903.39	1PULA	46.1	13747.8
<i>Trypanosoma brucei</i>	2AMHA	52.87	22994.83	1T9FA	38.98	20692.85
<i>Trypanosoma brucei</i>	2K9XA	40.33	12005.6	1TOVA	38.61	10827.16
<i>Trypanosoma cruzi</i>	1YZVA	44.88	22657.56	1MISA	34.33	12822.2

^a *L. donovani*.

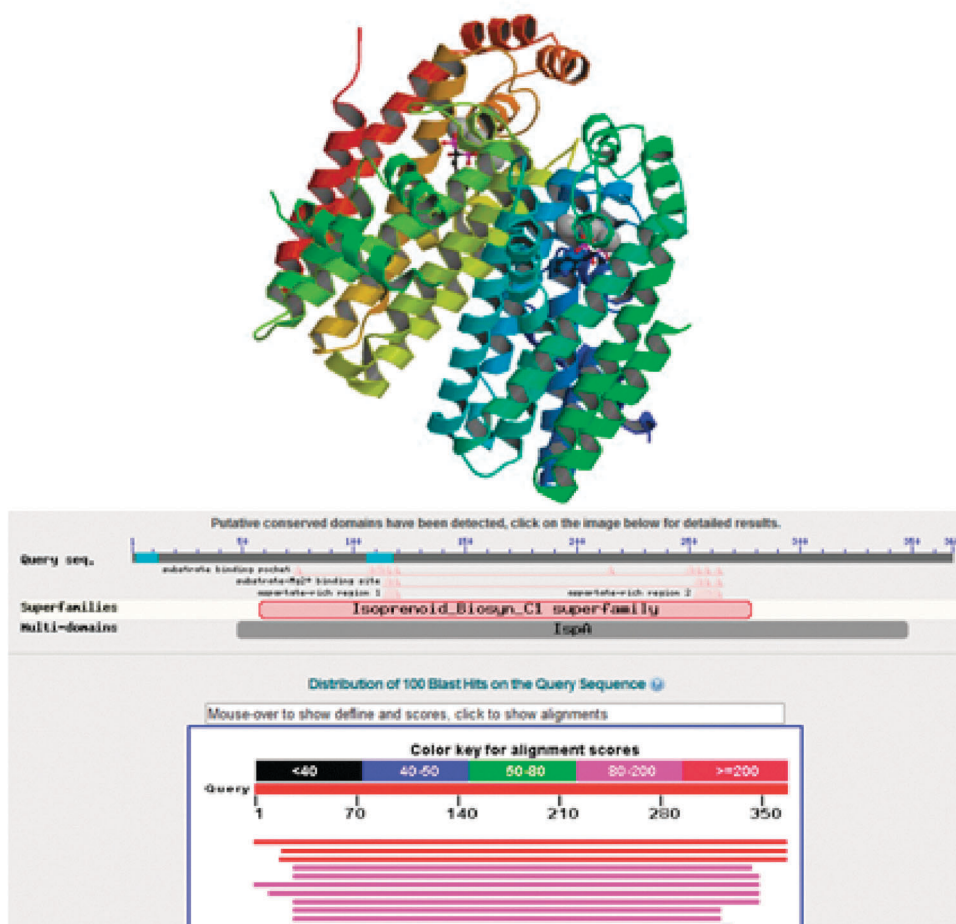


Fig. 5 BLAST analysis.

(a homodimer to be exact) with a total molecular weight of $M_w = 86659.06$. The importance of the study of proteins in this parasite is due to the fact that Cryptosporidiosis is a neglected disease, without a wholly effective drug. That is why Artz *et al.*¹²⁰ presented a study involving this protein in which they demonstrated that nitrogen-containing bisphosphonates (N-BPs) are capable of inhibiting *C. parvum* at low micromolar concentrations in infected MDCK cells.

Predictably, the mechanism of action is based on the inhibition of biosynthesis of isoprenoids but this target protein is unexpectedly a distinctive *C. parvum* enzyme that dubbed nonspecific polyprenyl pyrophosphate synthase (CpNPPPS). It is part of an isoprenoid pathway in *Cryptosporidium* distinctly different from other organisms. The proposed mechanism of action is corroborated by crystal structures of the enzyme with risedronate and zoledronate bonds showing how this enzyme's unique chain length determinant region enables it to accommodate larger substrates and products. N-BPs (such as pamidronate, alendronate, risedronate, ibandronate and zoledronate) seem to act as analogues of isoprenoid diphosphate lipids, thereby inhibiting FPP synthase, an enzyme in the mevalonate pathway.^{121–123} Interestingly, risedronate leads to an 88.9% inhibition of the rodent parasite *Plasmodium berghei*.¹²⁴ It may indicate that the prediction by LIBP-Pred as a potential drug target with LIBP function is correct and may break new ground to search for similar proteins in other parasites.

However, the protein is still reported as predicted with this putative enzyme action but function unknown. In any case, BLAST analysis also supports this idea by alignment, finding high homology between this protein and similar proteins in other organisms (see Fig. 5).

PDB mining of human proteome with LIBP-Pred

Considering that LIBPs/FABPs are very important cancer biomarkers in humans we decided to carry out a prediction of S(LIBP) values for all human proteins with unknown function in PDB. In Table 2 we summarized the most promising results found for human proteins, see also full results in ESI 1† (available online or upon author's request). We found a total of 168 proteins of the human proteome with unknown function and low sequence homology. After mining this dataset with LIBP-Pred we have predicted 15 out of these 168 proteins as LIBPs with $S(\text{LIBP}) > 50\%$. However only two proteins have a $S(\text{LIBP}) > 60$ and we have not found any protein with a higher value. The highest $S(\text{LIBP})$ values predicted for all human proteins studied with unknown function correspond to 2WM3 with $S(\text{LIBP}) = 62.66\%$. This is a statistically significant value (but not very high value indeed) of $S(\text{LIBP})$. Important clues that may support this prediction of 2WM3 by LIBP-Pred as a LIBP is the binding of this protein to both phosphate and glycerol separately, which are well-known components of

phospholipids. In any case, the protein header has an unknown function but also is bound to NADPH and is considered as an NmrA-like family domain-containing protein 1 in a public release to PDB. This theoretical result points out 2WM3 as a potential candidate for future experiments in the search of cancer biomarkers. For instance, human HSCARG has been annotated as a possible cancer related protein and also contains an NmrA-like domain.¹²⁵

Conclusions

The discovery of new LIBPs is a goal of great importance and several authors have presented interesting results. The present work has demonstrated that there is a strong linear relationship between electrostatic spectral moments calculated with a MARCH-INSIDE approach and the action of LIBPs. Consequently, using these parameters we can seek a linear QSAR useful to predict LIBPs. The online implementation of this model in the web server LIBP-Pred allows public researchers around the world to predict online new LIBPs free of cost. LIBP-Pred may be used to mine the PDB or to upload and predict custom 3D models of proteins with unknown structure generated with well-known servers as in the case of LOMETS. We have demonstrated the PDB mining option performing a predictive study of 2000+ proteins with unknown function looking for new Cancer Biomarkers in humans or drug targets in parasites.

Since user-friendly and publicly accessible web-servers represent the future direction of developing practically more useful predictors,¹²⁶ we have provided herein a web-server for the method presented in this paper at <http://miaja.tic.udc.es/Bio-AIMS/LIBPpred.php>.

Acknowledgements

Munteanu CR and González-Díaz H acknowledge the research programme Isidro Parga Pondal funded by Xunta de Galicia and the European Social Funds (ESF) for partial financial support. F. Prado-Prado acknowledges the research programme Angeles Albariño (funded by the same institutions) for partial financial support.

References

- J. Storch and L. McDermott, *J. Lipid Res.*, 2009, **50**(Suppl), S126–S131.
- E. Noiri, K. Doi, K. Negishi, T. Tanaka, Y. Hamasaki, T. Fujita, D. Portilla and T. Sugaya, *Am. J. Physiol.: Renal Physiol.*, 2009, **296**, F669–F679.
- T. L. Nickolas, J. Barasch and P. Devarajan, *Curr. Opin. Nephrol. Hypertens.*, 2008, **17**, 127–132.
- N. J. Evennett, M. S. Petrov, A. Mittal and J. A. Windsor, *World J. Surg.*, 2009, **33**, 1374–1383.
- E. Krusinova and T. Pelikanova, *Diabetes Res. Clin. Pract.*, 2008, **82**(Suppl 2), S127–S134.
- D. Abbanat, B. Morrow and K. Bush, *Curr. Opin. Pharmacol.*, 2008, **8**, 582–592.
- Y. Oka, A. Murata, J. Nishijima, T. Yasuda, N. Hiraoka, Y. Ohmachi, K. Kitagawa, T. Yasuda, H. Toda and N. Tanaka, *et al.*, *Cytokine*, 1992, **4**, 298–304.
- X. Y. Tang, S. Umemura, H. Tsukamoto, N. Kumaki, Y. Tokuda and R. Y. Osamura, *Pathol., Res. Pract.*, 2010, **206**, 98–101.
- C. Buhlmann, T. Borchers, M. Pollak and F. Spener, *Mol. Cell. Biochem.*, 1999, **199**, 41–48.
- Y. A. Yang, P. J. Morin, W. F. Han, T. Chen, D. M. Bornman, E. W. Gabrielson and E. S. Pizer, *Exp. Cell Res.*, 2003, **282**, 132–137.
- Z. Li, C. Huang, S. Bai, X. Pan, R. Zhou, Y. Wei and X. Zhao, *Int. J. Cancer*, 2008, **123**, 2377–2383.
- R. J. Morgan and I. Soltesz, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 6179–6184.
- M. L. De Santis, R. Hammamieh, R. Das and M. Jett, *J. Exp. Ther. Oncol.*, 2004, **4**, 91–100.
- R. Hammamieh, N. Chakraborty, R. Das and M. Jett, *J. Exp. Ther. Oncol.*, 2004, **4**, 195–202.
- L. McDermott, M. W. Kennedy, D. P. McManus, J. E. Bradley, A. Cooper and J. Storch, *Biochemistry*, 2002, **41**, 6706–6713.
- G. Zhu, *J. Eukaryotic Microbiol.*, 2004, **51**, 381–388.
- G. Greco, E. Novellino, I. Fiorini, V. Nacci, G. Campiani, S. M. Ciani, A. Garofalo, P. Bernasconi and T. Mennini, *J. Med. Chem.*, 1994, **37**, 4100–4108.
- M. Tendlar, C. A. Brito, M. M. Vilar, N. Serra-Freire, C. M. Diogo, M. S. Almeida, A. C. Delbem, J. F. Da Silva, W. Savino, R. C. Garratt, N. Katz and A. S. Simpson, *Proc. Natl. Acad. Sci. U. S. A.*, 1996, **93**, 269–273.
- F. Liu, S. J. Cui, W. Hu, Z. Feng, Z. Q. Wang and Z. G. Han, *Mol. Cell. Proteomics*, 2009, **8**, 1236–1251.
- L. McDermott, A. Cooper and M. W. Kennedy, *Mol. Cell. Biochem.*, 1999, **192**, 69–75.
- L. Kuang, M. L. Colgrave, N. H. Bagnall, M. R. Knox, M. Qian and G. Wiffels, *Mol. Biochem. Parasitol.*, 2009, **168**, 84–94.
- A. Hirasawa, T. Hara, S. Katsuma, T. Adachi and G. Tsujimoto, *Biol. Pharm. Bull.*, 2008, **31**, 1847–1851.
- J. P. Zbilut, A. Giuliani, A. Colosimo, J. C. Mitchell, M. Colafranceschi, N. Marwan, C. L. Webber, Jr. and V. N. Uversky, *J. Proteome Res.*, 2004, **3**, 1243–1253.
- B. Shen, J. Bai and M. Vihinen, *Protein Eng., Des. Sel.*, 2008, **21**, 37–44.
- J. Devillers and A. T. Balaban, *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach, The Netherlands, 1999.
- F. Torrens and G. Castellano, *Curr. Proteomics.*, 2009, **6**, 204–213.
- S. Thomas and D. Bonchev, *Hum. Genomics*, 2010, **4**, 353–360.
- D. Bonchev, S. Thomas, A. Apte and L. B. Kier, *SAR QSAR Environ. Res.*, 2010, **21**, 77–102.
- D. Bonchev and G. A. Buck, *J. Chem. Inf. Model.*, 2007, **47**, 909–917.
- L. B. Kier, D. Bonchev and G. A. Buck, *Chem. Biodiversity*, 2005, **2**, 233–243.
- D. Bonchev and D. H. Rouvray, *Complexity in Chemistry, Biology, and Ecology*, Springer Science + Business Media, Inc, New York, 2005.
- D. Bonchev, *Chem. Biodiversity*, 2004, **1**, 312–326.
- A. Duardo-Sanchez, G. Patlewicz and H. González-Díaz, *Curr. Bioinf.*, 2011, **6**, 53–70.
- P. Riera-Fernández, C. R. Munteanu, N. Pedreira-Souto, R. Martín-Romalde, A. Duardo-Sanchez and H. González-Díaz, *Curr. Bioinf.*, 2011, **6**, 94–121.
- H. Gonzalez-Diaz, *Curr. Pharm. Des.*, 2010, **16**, 2598–2600.
- H. Gonzalez-Diaz, F. Romaris, A. Duardo-Sanchez, L. G. Perez-Montoto, F. Prado-Prado, G. Patlewicz and F. M. Ubeira, *Curr. Pharm. Des.*, 2010, **16**, 2737–2764.
- R. Concu, G. Podda, F. M. Ubeira and H. Gonzalez-Diaz, *Curr. Pharm. Des.*, 2010, **16**, 2710–2723.
- J. Chen and B. Shen, *Curr. Proteomics*, 2009, **6**, 228–234.
- H. B. Shen and K. C. Chou, *Anal. Biochem.*, 2008, **373**, 386–388.
- H. B. Shen and K. C. Chou, *Protein Eng., Des. Sel.*, 2007, **20**, 561–567.
- K. C. Chou and H. B. Shen, *Biochem. Biophys. Res. Commun.*, 2007, **360**, 339–345.
- K. C. Chou and H. B. Shen, *Nat. Protocols*, 2008, **3**, 153–162.
- K. C. Chou, *J. Proteome Res.*, 2005, **4**, 1413–1418.
- K. C. Chou and D. W. Elrod, *J. Proteome Res.*, 2002, **1**, 429–433.
- K. C. Chou and D. W. Elrod, *J. Proteome Res.*, 2003, **2**, 183–190.
- K. C. Chou and H. B. Shen, *J. Proteome Res.*, 2006, **5**, 1888–1897.
- K. C. Chou and H. B. Shen, *J. Proteome Res.*, 2006, **5**, 3420–3428.
- K. C. Chou, *Curr. Proteomics*, 2009, **6**, 262–274.
- K. C. Chou, *J. Theor. Biol.*, 2011, **273**, 236–247.

- 50 L. Santana, E. Uriarte, H. González-Díaz, G. Zagotto, R. Soto-Otero and E. Mendez-Alvarez, *J. Med. Chem.*, 2006, **49**, 1149–1156.
- 51 H. González-Díaz, R. R. de Armas and R. Molina, *Bioinformatics*, 2003, **19**, 2079–2087.
- 52 G. Agüero-Chapin, J. Varona-Santos, G. A. de la Riva, A. Antunes, T. Gonzalez-Villa, E. Uriarte and H. Gonzalez-Diaz, *J. Proteome Res.*, 2009, **8**, 2122–2128.
- 53 H. González-Díaz, L. Saiz-Urra, R. Molina, L. Santana and E. Uriarte, *J. Proteome Res.*, 2007, **6**, 904–908.
- 54 R. Concu, M. A. Dea-Ayuela, L. G. Perez-Montoto, F. Bolas-Fernandez, F. J. Prado-Prado, G. Podda, E. Uriarte, F. M. Ubeira and H. Gonzalez-Diaz, *J. Proteome Res.*, 2009, **8**, 4372–4382.
- 55 L. Santana, H. Gonzalez-Diaz, E. Quezada, E. Uriarte, M. Yanez, D. Vina and F. Orallo, *J. Med. Chem.*, 2008, **51**, 6740–6751.
- 56 D. Vina, E. Uriarte, F. Orallo and H. Gonzalez-Diaz, *Mol. Pharmacol.*, 2009, **6**, 825–835.
- 57 H. Gonzalez-Diaz, F. Prado-Prado and F. M. Ubeira, *Curr. Top. Med. Chem.*, 2008, **8**, 1676–1690.
- 58 H. González-Díaz, Y. González-Díaz, L. Santana, F. M. Ubeira and E. Uriarte, *Proteomics*, 2008, **8**, 750–778.
- 59 H. González-Díaz, S. Vilar, L. Santana and E. Uriarte, *Curr. Top. Med. Chem.*, 2007, **7**, 1025–1039.
- 60 R. Concu, M. A. Dea-Ayuela, L. G. Perez-Montoto, F. J. Prado-Prado, E. Uriarte, F. Bolas-Fernandez, G. Podda, A. Pazos, C. R. Munteanu, F. M. Ubeira and H. Gonzalez-Diaz, *Biochim. Biophys. Acta*, 2009, **1794**, 1784–1794.
- 61 S. Vilar, H. Gonzalez-Diaz, L. Santana and E. Uriarte, *J. Theor. Biol.*, 2009, **261**, 449–458.
- 62 C. R. Munteanu, J. M. Vazquez, J. Dorado, A. P. Sierra, A. Sanchez-Gonzalez, F. J. Prado-Prado and H. Gonzalez-Diaz, *J. Proteome Res.*, 2009, **8**, 5219–5228.
- 63 R. Concu, M. A. Dea-Ayuela, L. G. Perez-Montoto, F. J. Prado-Prado, E. Uriarte, F. Bolas-Fernandez, G. Podda, A. Pazos, C. R. Munteanu, F. M. Ubeira and H. Gonzalez-Diaz, *Biochim. Biophys. Acta*, 2009, **1794**, 1784–1794.
- 64 R. Concu, M. A. Dea-Ayuela, L. G. Perez-Montoto, F. Bolas-Fernandez, F. J. Prado-Prado, G. Podda, E. Uriarte, F. M. Ubeira and H. Gonzalez-Diaz, *J. Proteome Res.*, 2009, **8**, 4372–4382.
- 65 Y. Rodriguez-Soca, C. R. Munteanu, J. Dorado, A. Pazos, F. J. Prado-Prado and H. Gonzalez-Diaz, *J. Proteome Res.*, 2010, **9**, 1182–1190.
- 66 C. R. M. Yamilet Rodriguez-Soca, J. Dorado, J. Rabuñal, A. Pazos and H. González-Díaz, *Polymer*, 2010, **51**, 264–273.
- 67 H. González-Díaz, A. Pérez-Bello and E. Uriarte, *Polymer*, 2005, **46**, 6461–6473.
- 68 L. Saiz-Urra, H. González-Díaz and E. Uriarte, *Bioorg. Med. Chem.*, 2005, **13**, 3641–3647.
- 69 H. González-Díaz, E. Uriarte and R. Ramos de Armas, *Bioorg. Med. Chem.*, 2005, **13**, 323–331.
- 70 R. Concu, G. Podda, E. Uriarte and H. Gonzalez-Diaz, *J. Comput. Chem.*, 2009, **30**, 1510–1520.
- 71 H. Gonzalez-Diaz, L. Saiz-Urra, R. Molina, Y. Gonzalez-Diaz and A. Sanchez-Gonzalez, *J. Comput. Chem.*, 2007, **28**, 1042–1048.
- 72 H. Gonzalez-Diaz, R. Molina and E. Uriarte, *FEBS Lett.*, 2005, **579**, 4297–4301.
- 73 R. Concu, G. Podda, E. Uriarte and H. Gonzalez-Diaz, *J. Comput. Chem.*, 2009, **30**, 1510–1520.
- 74 H. González-Díaz, Y. Pérez-Castillo, G. Podda and E. Uriarte, *J. Comput. Chem.*, 2007, **28**, 1990–1995.
- 75 StatSoft. Inc., 6.0 edn., 2002.
- 76 A. Speck-Planche, M. T. Scotti and V. de Paulo-Emerenciano, *Curr. Pharm. Des.*, 2010, **16**, 2656–2665.
- 77 A. Speck-Planche and M. N. D. S. Cordeiro, *Curr. Bioinf.*, 2011, **6**, 81–93.
- 78 A. Speck-Planche, M. T. Scotti, V. P. Emerenciano, A. García-López, E. Molina-Pérez and E. Uriarte, *J. Comput. Chem.*, 2010, **31**, 882–894.
- 79 A. Speck-Planche, M. T. Scotti, A. García-López, V. P. Emerenciano, E. Molina-Pérez and E. Uriarte, *Mol. Diversity*, 2009, **13**, 445–458.
- 80 A. Speck-Planche, L. Guilarte-Montero, R. Yera-Bueno, J. A. Rojas-Vargas, A. Garcia-Lopez, E. Uriarte and E. Molina-Perez, *Pest Manage. Sci.*, 2011, **67**, 438–445.
- 81 A. Speck-Planche, V. V. Kleandrova and J. A. Rojas-Vargas, *Mol. Diversity*, 2011, **15**, 901–909.
- 82 A. Speck-Planche, V. V. Kleandrova, F. Luan and M. N. Cordeiro, *Bioorg. Med. Chem.*, 2011, **19**, 6239–6244.
- 83 G. M. Casanola-Martin, M. T. Khan, Y. Marrero-Ponce, A. Ather, M. N. Sultankhodzhaev and F. Torrens, *Bioorg. Med. Chem. Lett.*, 2006, **16**, 324–330.
- 84 G. M. Casanola-Martin, Y. Marrero-Ponce, M. T. Khan, A. Ather, K. M. Khan, F. Torrens and R. Rotondo, *Eur. J. Med. Chem.*, 2007, **42**, 1370–1381.
- 85 G. M. Casanola-Martin, Y. Marrero-Ponce, M. T. Khan, A. Ather, S. Sultan, F. Torrens and R. Rotondo, *Bioorg. Med. Chem.*, 2007, **15**, 1483–1503.
- 86 G. M. Casanola-Martin, Y. Marrero-Ponce, M. Tareq Hassan Khan, F. Torrens, F. Perez-Gimenez and A. Rescigno, *J. Biomol. Screening*, 2008, **13**, 1014–1024.
- 87 Y. Marrero-Ponce, R. Medina-Marrero, A. E. Castro, R. Ramos de Armas, H. González-Díaz, V. Romero-Zaldivar and F. Torrens, *Molecules*, 2004, **9**, 1124–1147.
- 88 R. Ramos de Armas, H. Gonzalez Diaz, R. Molina and E. Uriarte, *Proteins: Struct., Funct., Genet.*, 2004, **56**, 715–723.
- 89 R. Ramos de Armas, H. González-Díaz, R. Molina, M. Perez Gonzalez and E. Uriarte, *Bioorg. Med. Chem.*, 2004, **12**, 4815–4822.
- 90 R. Ramos de Armas, H. González-Díaz, R. Molina and E. Uriarte, *Biopolymers*, 2005, **77**, 247–256.
- 91 A. Speck-Planche, M. T. Scotti and V. de Paulo-Emerenciano, *Curr. Pharm. Des.*, 2010, **16**, 2656–2665.
- 92 K. C. Chou and C. T. Zhang, *Crit. Rev. Biochem. Mol. Biol.*, 1995, **30**, 275–349.
- 93 K. C. Chou and H. B. Shen, *Nat. Sci.*, 2010, **2**, 1090–1103 (openly accessible at <http://www.scirp.org/journal/NS/>).
- 94 K. C. Chou, Z. C. Wu and X. Xiao, *Mol. BioSyst.*, 2012, DOI: 10.1039/C1MB05420A.
- 95 M. Esmaili, H. Mohabatkar and S. Mohsenzadeh, *J. Theor. Biol.*, 2010, **263**, 203–209.
- 96 D. N. Georgiou, T. E. Karakasidis, J. J. Nieto and A. Torres, *J. Theor. Biol.*, 2009, **257**, 17–26.
- 97 Q. Gu, Y. S. Ding and T. L. Zhang, *Protein Pept. Lett.*, 2010, **17**, 559–567.
- 98 H. Mohabatkar, *Protein Pept. Lett.*, 2010, **17**, 1207–1214.
- 99 H. Mohabatkar, M. Mohammad Beigi and A. Esmaili, *J. Theor. Biol.*, 2011, **281**, 18–23.
- 100 L. Yu, Y. Guo, Y. Li, G. Li, M. Li, J. Luo, W. Xiong and W. Qin, *J. Theor. Biol.*, 2010, **267**, 1–6.
- 101 J. D. Qiu, J. H. Huang, S. P. Shi and R. P. Liang, *Protein Pept. Lett.*, 2010, **17**, 715–722.
- 102 K. C. Chou, Z. C. Wu and X. Xiao, *PLoS One*, 2011, **6**, e18258.
- 103 X. Xiao, P. Wang and K. C. Chou, *Mol. Diversity*, 2011, **15**, 149–155.
- 104 V. A. Ivanisenko, S. S. Pintus, D. A. Grigorovich and N. A. Kolchanov, *Nucleic Acids Res.*, 2005, **33**, D183–D187.
- 105 P. D. Dobson and A. J. Doig, *J. Mol. Biol.*, 2003, **330**, 771–783.
- 106 Z. C. Wu, X. Xiao and K. C. Chou, *Mol. BioSyst.*, 2011, **7**, 3287–3297.
- 107 X. Xiao, Z. C. Wu and K. C. Chou, *J. Theor. Biol.*, 2011, **284**, 42–51.
- 108 X. Xiao, Z. C. Wu and K. C. Chou, *PLoS One*, 2011, **6**(6), e20592.
- 109 M. Perez Gonzalez and A. Morales Helguera, *J. Comput.-Aided Mol. Des.*, 2003, **17**, 665–672.
- 110 Y. Marrero-Ponce, R. Medina-Marrero, F. Torrens, Y. Martinez, V. Romero-Zaldivar and E. A. Castro, *Bioorg. Med. Chem.*, 2005, **13**, 2881–2899.
- 111 Y. Marrero-Ponce, A. Montero-Torres, C. R. Zaldivar, M. I. Veitia, M. M. Perez and R. N. Sanchez, *Bioorg. Med. Chem.*, 2005, **13**, 1293–1304.
- 112 H. González-Díaz, A. Sanchez-Gonzalez and Y. Gonzalez-Diaz, *J. Inorg. Biochem.*, 2006, **100**, 1290–1297.
- 113 StatSoft.Inc., 6.0 edn., 2002.
- 114 H. Van Waterbeemd, in *Method and Principles in Medicinal Chemistry*, ed. R. Manhnhold, P. Krosggaard-Larsen, H. Timmerman and H. Van Waterbeemd, Wiley-VCH, New York, 1995, vol. 2, pp. 283–293.
- 115 S. Wu and Y. Zhang, *Nucleic Acids Res.*, 2007, **35**, 3375–3382.
- 116 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 and F. M. Ubeira, *J. Proteome Res.*, 2011, **10**, 1698–1718.
- 117 E. Estrada, E. Uriarte, A. Montero, M. Teijeira, L. Santana and E. De Clercq, *J. Med. Chem.*, 2000, **43**, 1975–1985.
- 118 D. Ramel, F. Lagarrigue, V. Pons, J. Mounier, S. Dupuis-Coronas, G. Chicanne, P. J. Sansonetti, F. Gaits-Iacovoni, H. Tronchere and B. Payrastra, *Sci. Signaling*, 2011, **4**, ra61.
- 119 E. Mikiciuk-Olasik, E. Zurek, R. Mikolajczak, E. Zakrzewska and K. Blaszcak-Swiatkiewicz, *Nucl. Med. Rev. Cent. East. Eur.*, 2000, **3**, 149–152.
- 120 J. D. Artz, J. E. Dunford, M. J. Arrowood, A. Dong, M. Chruszcz, K. L. Kavanagh, W. Minor, R. G. Russell, F. H. Ebetino, U. Oppermann and R. Hui, *Chem. Biol.*, 2008, **15**, 1296–1306.
- 121 A. A. Reszka and G. A. Rodan, *Mini-Rev. Med. Chem.*, 2004, **4**, 711–719.
- 122 A. A. Reszka and G. A. Rodan, *Curr. Rheumatol. Rep.*, 2003, **5**, 65–74.
- 123 G. A. Rodan and A. A. Reszka, *Curr. Mol. Med.*, 2002, **2**, 571–577.
- 124 F. M. Jordao, A. Y. Saito, D. C. Miguel, V. de Jesus Peres, E. A. Kimura and A. M. Katzin, *Antimicrob. Agents Chemother.*, 2011, **55**, 2026–2031.
- 125 X. Dai, X. Gu, M. Luo and X. Zheng, *Protein Pept. Lett.*, 2006, **13**, 955–957.
- 126 K. C. Chou and H. B. Shen, *Nat. Sci.*, 2009, **2**, 63–92.