UNIVERSIDADE DA CORUÑA

Facultad de Informática

Departamento de Tecnologías de la
Información
y las Comunicaciones

# Herramientas Informáticas y de Inteligencia Artificial para el Meta-Análisis En la Frontera entre la Bioinformática y las Ciencias Jurídicas

**Tesis Doctoral**

Doctoranda

**ALIUSKA DUARDO SANCHEZ**

Directores

Prof. Dr. Alejandro Pazos Sierra

Prof. Dr. Humberto González Díaz

A Coruña, Febrero **2014**

**Prof. Dr. D. Alejandro Pazos Sierra**, Catedrático del Departamento de Tecnologías de la Información y las Comunicaciones, Facultad de Informática, Universidade da Coruña, UDC.

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

**CERTIFICAN:**

Que la memoria titulada "**HERRAMIENTAS INFORMÁTICAS Y DE INTELIGENCIA ARTIFICIAL PARA EL META-ANÁLISIS EN LA FRONTERA ENTRE LA BIOINFORMÁTICA Y LAS CIENCIAS JURIÍDICAS**" presentada por **D. ALIUSKA DUARDO SANCHEZ**, ha sido realizada bajo nuestra dirección, en el Departamento de Tecnologías de la Información y las Comunicaciones, Facultad de Informática, Universidade da Coruña, UDC y en el Departamento de Química Orgánica II, Facultad de Ciencia y Tecnología, Universidad del País Vasco UPV/EHU.

Considerando que el trabajo constituye tema de Tesis Doctoral, se autoriza su presentación en la Universidade da Coruña.

Y para que conste, se expide el presente certificado en A Coruña, a 3 de Febrero de dos mil catorce.

_____

Firmado Prof. Dr. Alejandro Pazos Sierra

_____

Firmado Prof. Dr. Humberto González Díaz

*A Juan, mi hijo.*

*A Juan, mi abuelo.*

# AGRADECIMIENTOS

Agradezco profundamente la ayuda prestada por los directores de esta tesis Alejandro Pazos Sierra y Humberto González Díaz, así como a sus colaboradores de la Universidade da Coruña, especialmente al Dr. Cristian Robert Munteanu del grupo de Redes de Neuronas Artificiales y Sistemas Adaptativos RNASA-IMEDIR.

Me gustaría extender estos agradecimientos a todos los amigos que en España me han apoyado tanto humana como profesionalmente; y que han hecho posible, de un modo u otro, este trabajo. Entre ellos, muy significativamente a: Antonio, Begoña, Eugenio, Lourdes, Cristian, Mari, Susana, Luisa, Xavi, Fran, Paola, Chachi, Dayamí, Ricci, Oana, Yanela, Rosi, Yaque, Sergio, Alejandro y Valeria.

Quiero hacer una mención especial a mi familia, sin cuyo esfuerzo y apoyo no habría sido posible llegar hasta aquí, aunque la vida nos haya dispersado por el mundo. Mi hermano y mi primo Yoel, en Houston. Mis primos Berkis, Enrique, Mari, Luis, y la tía Clara, en Madrid. A mi cuñada Yenny, en Brasil. A mi familia en Cuba: Julio, Ana y Humberto, Marquitos y Humbertico, en Santa Clara. Por último y muy especialmente, al resto de mi familia cubana en Sancti Spiritus, La Habana, Güinía, y Manicaragüa.

¡A todos muchas Gracias!

**Resumen**

Los modelos computacionales, conocidos por su acrónimo en idioma Inglés como QSPR (Quantitative Structure-Property Relationships) pueden usarse para predecir propiedades de sistemas complejos. Estas predicciones representan una aplicación importante de las Tecnologías de la Información y la Comunicación (TICs). La mayor relevancia es debido a la reducción de costes de medición experimental en términos de tiempo, recursos humanos, recursos materiales, y/o el uso de animales de laboratorio en ciencias bio-moleculares, técnicas, sociales y/o jurídicas.

Las Redes Neuronales Artificiales (ANNs) son una de las herramientas informáticas más poderosas para buscar modelos QSPR. Para ello, las ANNs pueden usar como variables de entrada (input) parámetros numéricos que cuantifiquen información sobre la estructura del sistema. Los parámetros conocidos como Índices Topológicos (TIs) se encuentran entre los más versátiles.

Los TIs se calculan en Teoría de Grafos a partir de la representación de cualquier sistema como una red de nodos interconectados; desde moléculas a redes biológicas, tecnológicas, y sociales. Esta tesis tiene como primer objetivo realizar una revisión y/o introducir nuevos TIs y software de cálculo de TIs útiles como inputs de ANNs para el desarrollo de modelos QSPR de redes bio-moleculares, biológicas, tecnológico-económicas y socio-jurídicas. En ellas, por una parte, los nodos representan bio-moléculas, organismos, poblaciones, leyes tributarias o concausas de delitos. Por otra parte, en la interacción TICs-Ciencias Biomoleculares-Derecho se hace necesario un marco de seguridad jurídica que permita el adecuado desarrollo de las TICs y sus aplicaciones en Ciencias Bio-moleculares. Por eso, el segundo objetivo de esta tesis es revisar el marco jurídico-legal de protección de los modelos QSAR/QSPR de sistemas moleculares.

El presente trabajo de investigación pretende demostrar la utilidad de estos modelos para predecir características y propiedades de estos sistemas complejos.

**Resumo**

Os modelos de ordenador coñecidos pola súas iniciais en inglés QSPR (Quantitative Structure-Property Relationships) poden prever as propiedades de sistemas complexos e reducir os custos experimentais en termos de tempo, recursos humanos, materiais e/ou o uso de animais de laboratorio nas ciencias biomoleculares, técnicas, e sociais.

As Redes Neurais Artificiais (ANNs) son unha das ferramentas máis poderosas para buscar modelos QSPR. Para iso, as ANNs poden facer uso, coma variables de entrada (input), dos parámetros numéricos da estrutura do sistema chamados Índices Topolóxicos (TIs).

Os TI calcúlanse na teoría dos grafos a partir da representación do sistema coma unha rede de nós conectados, incluíndo tanto moléculas coma redes sociais e tecnolóxicas. Esta tese ten como obxectivo principal revisar e/ou desenvolver novos TIs, programas de cálculo de TIs, e/ou modelos QSPR facendo uso de ANNs para predicir redes bio-moleculares, biolóxicas, económicas, e sociais ou xurídicas onde os nós representan moléculas biolóxicas, organismos, poboacións, ou as leis fiscais ou as concausas dun delito. Ademais, a interacción das TIC con as ciencias biolóxicas e xurídicas necesita dun marco de seguridade xurídica que permita o bo desenvolvemento das TIC e as súas aplicacións en Ciencias Biomoleculares. Polo tanto, o segundo obxectivo desta tese é analizar o marco xurídico e legal de protección dos modelos QSPR.

O presente traballo de investigación pretende demostrar a utilidade destes modelos para predicir características e propiedades destes sistemas complexos.

## Abstract

QSPR (Quantitative Structure-Property Relationships) computer models can predict properties of complex systems reducing experimental costs in terms of time, human resources, material resources, and/or the use of laboratory animals in bio-molecular, technical, and/or social sciences.

Artificial Neural Networks (ANNs) are one of the most powerful tools to search QSPR models. For this, the ANNs may use as input variables numerical parameters of the system structure called Topological Indices (TIs).

The TIs are calculated in Graph Theory from a representation of any system as a network of interconnected nodes, including molecules or social and technological networks. The first aim of this thesis is to review and/or develop new TIs, TIs calculation software, and QSPR models using ANNs to predict bio-molecular, biological, commercial, social, and legal networks where nodes represent bio-molecules, organisms, populations, products, tax laws, or criminal causes. Moreover, the interaction of ICTs with Biomolecular and law Sciences needs a legal security framework that allows the proper development of ICTs and their applications in Bio-molecular Sciences. Therefore, the second objective of this thesis is to review the legal framework and legal protection of QSPR techniques.

The present work of investigation tries to demonstrate the usefulness of these models to predict characteristics and properties of these complex systems.

**ÍNDICE**

# I. INTRODUCCIÓN Y OBJETIVOS

# 1. Introducción

*Un camino de ida y vuelta entre las TICs y las Ciencias Jurídicas y Bio-Moleculares*

**A. Duardo-Sanchez**, G. Patlewicz, and H. González-Díaz. Network Topological Indices from Chem-Bioinformatics to Legal Sciences and back. *Current Bioinformatics*, **2011**, 6(1), 53-70.

En la actualidad, es difícil encontrar un área, ya sea científica o no, en la que no exista una aplicación importante de las TICs. Incluso los sistemas jurídicos, en sí mismos, han sido revolucionados por el uso de las llamadas TICs. En este sentido, se hace necesario un marco jurídico que proporcione una base segura y pertinente para el desarrollo de dichas tecnologías. En particular, la Quimio-informática, la Bioinformática, la Biología computacional, y otras áreas afines dentro del campo de las Biociencias, están entre las que requieren más atención de los operadores legales debido, entre otras cosas, a su novedad. Todo ello hace preciso el uso de las TICs en el estudio de procesos biológicos y sistemas bio-moleculares complejos. Pudiera decirse que tales áreas constituyen esencialmente la interrelación de las TICs con las Ciencias de la vida. No hay que olvidar que, en los últimos 20 años, más del 70% del descubrimiento científico se ha realizado en la intersección de estas dos áreas de la ciencia, creándose grandes lagunas legales al respecto. El Derecho tiene y debe tener un impacto importante en la regulación de las TICs aplicadas a problemas quimio-informáticos y bio-informáticos. De hecho, la relación entre ellos engloba una gama amplia de asuntos diferentes que incluyen propiedad intelectual (por ejemplo: marcas, secreto industrial, propiedad intelectual de software,…), licencias, patentes legislación, control y desarrollo de productos, Así como asuntos jurídicos corporativos, incluso cuestiones ético-legales. Todo esto involucra el uso de herramientas y técnicas de tres disciplinas separadas; Ciencias Biomoleculares (la fuente de los datos que deben ser analizados), Ciencias de la Computación e Inteligencia Artificial (proponen las herramientas TICs, incluyendo el *hardware* para

hacer cálculos, el *software* y algoritmos de análisis de datos). El desarrollo *software* para Bioinformática es un reto para el presente y un negocio para el futuro; facilitando la creación de nuevas empresas de TICs como soporte a las Ciencias Biomoleculares.

Los modelos Quimio-informáticos y Bio-informáticos que describen relaciones estructura-propiedad de sistemas biológicos complejos (*complex biosystems*) pueden jugar una función importante para reducir costes en lo concerniente a: plazos de tiempo, recursos humanos, recursos materiales, así como de sustitución de animales de laboratorio en Ciencias Biomédicas. Muchos de estos modelos son, en esencia, Relaciones Cuantitativas Estructura-Actividad o Propiedad conocidos por sus siglas en Inglés: *Quantitative Structure-Activity or Property Relationships (QSAR/QSPR)*. En otras palabras, los modelos informáticos QSPRs conectan la información sobre la estructura de un sistema con información sobre propiedades externas de estos sistemas que no son evidentes después de una inspección visual directa de la estructura. En particular, los modelos QSARs conectan la información sobre la estructura química de fármacos y dianas moleculares (proteína, gen, ARN, microorganismos, tejidos, enfermedades…, etc.) con la actividad biológica del fármaco sobre sus posibles dianas. Muchos de estos modelos QSAR están basados en el uso de parámetros químico-informáticos estructurales. Tales parámetros son series numéricas que codifican información estructural para predecir correlaciones entre la estructura molecular y propiedades biológicas. Existen diversas herramientas informáticas para encontrar un modelo QSPR que relacione los TIs de un sistema con sus propiedades. Una de las herramientas más usadas con este objeto son las Redes Neuronales Artificiales (ANNs, por sus siglas en inglés). Una actividad intensiva en este campo ha incitado a muchos autores a publicar revisiones recientes y/o editar números especiales en revistas de alto impacto relacionadas con estos temas. Por ejemplo, González-Díaz H, y col., han editado varios números especiales con artículos de grupos de diferentes países que revisan el uso de estos métodos en las revistas: *Current Topics in Medicinal Chemistry* en 2008, *Current Proteomics* en 2009, *Current Durgs Metabolism*, y *Current Pharmaceutical Design* en 2010.

En paralelo, las herramientas de análisis basadas en la Teoría de Grafos y Redes Complejas se están expandiendo a nuevos campos de aplicación. Estas herramientas usadas en las TICs permiten interconectar información a niveles estructurales de la materia muy diferentes desde moléculas a redes macroscópicas. Constituyen ejemplos de lo que se acaba de decir: los genomas, las redes de interacción proteína-proteína, las redes complejas metabólicas, redes neuronales biológicas y artificiales, redes sociales y tecnológicas (o ambas). Entre estas últimas se destacan las redes de transmisión de enfermedades, redes de distribución eléctrica, redes de telefonía inalámbricas, redes de comercio internacional, Internet y, soportada por ella, la *world wide web* (*www*), *facebook, twitter, etc*. En todos estos casos, se pueden calcular un tipo de parámetros llamados Índices Topológicos (TIs) que describen numéricamente los patrones de conectividad existentes entre los nodos o actores en una red (representada como un grafo matemático). Los TIs pueden ser muy útiles como entradas (variables *input*) para desarrollo de modelos QSPR en escalas múltiples (multiescala). En particular, el caso de las relaciones entre actores sociales, así como las relaciones entre actores en diferentes niveles de análisis (como personas y grupos) se ha extendido como tema de investigación.

El análisis de redes sociales, *Social Networks Analysis (SNA)*, proporciona una aproximación común para todas aquellas disciplinas que implican un estudio de la estructura social susceptible de representación en forma de redes. El concepto de estructura social es comúnmente utilizado en sociología y en la teoría social. A pesar de que no hay acuerdo entre los teóricos, puede referirse a un tipo específico de relación entre las entidades o los grupos que también pueden evolucionar a patrones duraderos de comportamiento y relación dentro de una sociedad, o normas e instituciones sociales que acontecen dentro de sistemas sociales. Para una revisión completa sobre SNA, se puede consultar la revisión llevada a cabo por Newman en su artículo: *La Estructura y Función de Redes Complejas*. En todo caso, si se toma en consideración que una red es un conjunto de elementos (nodos) con conexiones entre ellos llamadas aristas del grafo, se pueden representar las relaciones sociales en forma

de redes, donde los nodos pueden ser los actores individuales dentro de las redes y las aristas las relaciones entre dichos actores. De hecho, la utilización de SNA no es nada nuevo en estudios de ciencias sociales. Ya en los tempranos años 30 del siglo pasado, los sociólogos usaban redes sociales para estudiar relaciones de amistad entre niños en edad escolar. Desde entonces, SNA ha sido aplicado al estudio de las interrelaciones entre miembros de una misma familia, interacción empresarial, o patrones de relaciones sexuales. A pesar de que la aproximación SNA está muy difundida en las ciencias sociales, su aplicación en el en Ciencias Jurídicas es escasa. Las herramientas de redes complejas y las metodologías TICs pueden ser útiles para ilustrar la interrelación entre los tipos diferentes de leyes. También pueden ayudar a predecir y entender las consecuencias de la aprobación de nuevas leyes en la sociedad y su efectividad. Tomando todos los aspectos anteriores en consideración, esta tesis profundizará en ambos sentidos: el uso de Ciencias Jurídicas para propiciar un marco jurídico adecuado de protección de los modelos QSAR/QSPR de sistemas moleculares y la aplicación de modelos QSPR para estudiar sistemas jurídicos *per se*. Por ello, los objetivos de la tesis y la hipótesis de trabajo de partida son los siguientes:

**Objetivos:**

1- Revisar las opciones legales disponibles para proteger los programas de computador (software) usados para desarrollo de modelos QSPR/QSAR, la aceptación y tratamiento legal de resultados científicos obtenidos con este tipo de software, así como algunos asuntos financiero-tributarios específicos en el campo de los programas del ordenador.

2- Revisar la representación de los sistemas jurídicos que utilizan redes complejas, definir nuevos TIs para la descripción de redes bio-moleculares y jurídico-sociales, y desarrollar nuevos modelos QSPR con ANNs para pronosticar fenómenos bio-moleculares, jurídicos, y otros.

**Hipótesis de trabajo:**

La hipótesis de trabajo de partida se formula como la posibilidad de demostrar la utilidad y adecuación de los modelos QSAR / QSPR de sistemas moleculares y de técnicas/herramientas de informática convencional y de inteligencia artificial para predecir características y propiedades de sistemas complejos. Así, se aplicarán en ámbitos complejos tan diversos como son el biomolecular, el tecnológico y el jurídico social.

**Bibliografía consultada**


Abercrombie, N.; Hill, S.; Turner, B.S. Social structure. In *The Penguin Dictionary of Sociology*, 4th ed.; Penguin: London, **2000**.

Bornholdt, S.; Schuster, H.G. Handbook of Grafos and Complex Networks: From the Genome to the Internet. WILEY-VCH GmbH & CO. KGa.: Wheinheim 2003.

Breiger, R. The Analysis of Social Networks. In *Handbook of Data Analysis*, Hardy, M.; Bryman, A., Eds. Sage Publications: London, **2004**; pp 505-26.

Chen, J.; Shen, B. Computational Analysis of Amino Acid Mutation: a Proteome Wide Perspective. *Curr Proteomics* **2009**; 6: 228-34.

Chou, K.C. Grafoic rule for drug metabolism systems. *Curr Drug Metab* **2010**; 11: 369-78.

Chou, K.C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr Proteomics* **2009**; 6: 262-74.

Concu, R.; Dea-Ayuela, M.A.; Perez-Montoto, L.G.; Prado-Prado, F.J.; Uriarte, E.; Bolas-Fernandez, F.; Podda, G.; Pazos, A.; Munteanu, C.R.; Ubeira, F.M.; Gonzalez-Diaz, H. 3D entropy and moments prediction of enzyme classes and experimental-theoretic study of peptide fingerprints in Leishmania parasites. *Biochimica et Biophysica Acta* **2009**; 1794: 1784-94.

Concu, R.; Podda, G.; Ubeira, F.M.; Gonzalez-Diaz, H. Review of QSAR models for enzyme classes of drug targets: Theoretical background and applications in parasites, hosts, and other organisms. *Current Pharmaceutical Design* **2010**; 16: 2710-23.

Craig, C. Social Structure. In *Dictionary of the Social Sciences*, Oxford University Press: Oxford, **2002**.

Duardo-Sanchez, A.; Gonzalez-Diaz H. Legal issues for chem-bioinformatics models. *Frontiers in Biosciences* (Elite Ed). **2013**; 5: 361-374.

Duardo-Sanchez, A.; Patlewicz, G.; and González-Díaz, H. Network Topological Indices from Chem-Bioinformatics to Legal Sciences and back. *Current Bioinformatics*, **2011**, 6(1), 53-70.

Duardo-Sanchez, A.; Patlewicz, G.; Lopez-Diaz, A. Current topics on software use in medicinal chemistry: intellectual property, taxes, and regulatory issues. *Curr Top Med Chem* **2008**; 8: 1666-75.

Estrada, E.; Molina, E.; Nodarse, D.; Uriarte, E. Structural contributions of substrates to their binding to P-Glycoprotein. A TOPS-MODE approach. *Current Pharmaceutical Design* **2010**; 16: 2676-709.

Garcia, I.; Diop, Y.F.; Gomez, G. QSAR & complex network study of the HMGR inhibitors structural diversity. *Curr Drug Metab* **2010**; 11: 307-14.

Garcia, I.; Fall, Y.; Gomez, G. QSAR, docking, and CoMFA studies of GSK3 inhibitors. *Current Pharmaceutical Design* **2010**; 16: 2666-75.

Giuliani, A.; Di Paola, L.; Setola, R. Proteins as Networks: A Mesoscopic Approach Using Haemoglobin Molecule as Case Study. *Curr Proteomics* **2009**; 6: 235-45.

Gonzalez-Diaz, H. Network topological indices, drug metabolism, and distribution. *Curr Drug Metab* **2010**; 11: 283-4.

Gonzalez-Diaz, H. QSAR and complex networks in pharmaceutical design, microbiology, parasitology, toxicology, cancer, and neurosciences. *Current Pharmaceutical Design* **2010**; 16: 2598-600.

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

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

Gonzalez-Diaz, H.; Prado-Prado, F.; Ubeira, F.M. Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. *Curr Top Med Chem* **2008**; 8: 1676-90.

Gonzalez-Diaz, H.; Romaris, F.; Duardo-Sanchez, A.; Perez-Montoto, L.G.; Prado-Prado, F.; Patlewicz, G.; Ubeira, F.M. Predicting drugs and proteins in parasite infections with topological indices of complex networks: theoretical backgrounds, applications, and legal issues. *Current Pharmaceutical Design* **2010**; 16: 2737-64.

Helguera, A.M.; Combes, R.D.; Gonzalez, M.P.; Cordeiro, M.N. Applications of 2D descriptores in drug design: a DRAGON tale. *Curr Top Med Chem* **2008**; 8: 1628-55.

Ivanciuc, O. Machine learning Quantitative Structure-Activity Relationships (QSAR) for peptides binding to Human Amphiphysin-1 SH3 domain. *Curr Proteomics* **2009**; 4: 289-302.

Ivanciuc, O. Weka machine learning for predicting the phospholipidosis inducing potential. *Curr Top Med Chem* **2008**; 8: 1691-709.

Khan, M.T. Predictions of the ADMET properties of candidate drug molecules utilizing different QSAR/QSPR modelling approaches. *Curr Drug Metab* **2010**; 11: 285-95.

Marrero-Ponce, Y.; Casanola-Martin, G.M.; Khan, M.T.; Torrens, F.; Rescigno, A.; Abad, C. Ligand-based computer-aided discovery of tyrosinase inhibitors.

Applications of the TOMOCOMD-CARDD method to the elucidation of new compounds. *Current Pharmaceutical Design* **2010**; 16: 2601-24.

Martinez-Romero, M.; Vazquez-Naya, J.M.; Rabunal, J.R.; Pita-Fernandez, S.; Macenlle, R.; Castro-Alvarino, J.; Lopez-Roses, L.; Ulla, J.L.; Martinez-Calvo, A.V.; Vazquez, S.; Pereira, J.; Porto-Pazos, A.B.; Dorado, J.; Pazos, A.; Munteanu, C.R. Artificial intelligence técnicas for colorectal cancer drug metabolism: ontology and complex network. *Curr Drug Metab* **2010**; 11: 347-68.

Mrabet, Y.; Semmar, N. Mathematical methods to analysis of topology, functional variability and evolution of metabolic systems based on different decomposition concepts. *Curr Drug Metab* **2010**; 11: 315-41.

Munteanu, C.R.; Fernandez-Blanco, E.; Seoane, J.A.; Izquierdo-Novo, P.; Rodriguez-Fernandez, J.A.; Prieto-Gonzalez, J.M.; Rabunal, J.R.; Pazos, A. Drug discovery and design for complex diseases through QSAR computational methods. *Current Pharmaceutical Design* **2010**; 16: 2640-55.

Newman, M. The Structure and Function of Complex Networks. *SIAM Review* **2003**; 56: 167-256.

Roy, K.; Ghosh, G. Exploring QSARs with Extended Topochemical Atom (ETA) indices for modeling chemical and drug toxicity. *Current Pharmaceutical Design* **2010**; 16: 2625-39.

Speck-Planche, A.; Scotti, M.T.; de Paulo-Emerenciano, V. Current pharmaceutical design of antituberculosis drugs: future perspectives. *Current Pharmaceutical Design* **2010**; 16: 2656-65.

Torrens, F.; Castellano, G. Topological Charge-Transfer Indices: From Small Molecules to Proteins. *Curr Proteomics* **2009**: 204-13.

Vazquez-Naya, J.M.; Martinez-Romero, M.; Porto-Pazos, A.B.; Novoa, F.; Valladares-Ayerbes, M.; Pereira, J.; Munteanu, C.R.; Dorado, J. Ontologies of drug discovery and design for neurology, cardiology and oncology. *Current Pharmaceutical Design* **2010**; 16: 2724-36.

Vilar, S.; Cozza, G.; Moro, S. Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. *Curr Top Med Chem* **2008**; 8: 1555-72.

Vilar, S.; Gonzalez-Diaz, H.; Santana, L.; Uriarte, E. A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer. *Journal of Theoretical Biology* **2009**; 261: 449-58.

Wang, J.F.; Chou, K.C. Molecular modeling of cytochrome P450 and drug metabolism. *Curr Drug Metab* **2010**; 11: 342-6.

Wang, J.F.; Wei, D.Q.; Chou, K.C. Drug candidates from traditional chinese medicines. *Curr Top Med Chem* **2008**; 8: 1656-65.

Wang, J.F.; Wei, D.Q.; Chou, K.C. Pharmacogenomics and personalized use of drugs. *Curr Top Med Chem* **2008**; 8: 1573-9.

Wellman, B.; Berkowitz, S.D. Social Structures: A Network Approach. Cambridge University Press: Cambridge 1988.

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.

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

# II. FUNDAMENTOS TEÓRICOS

## II.1. Redes, Índices Topológicos, y modelos QSAR/QSPR

Los grafos se están mostrando muy útiles para representar la estructura de los fármacos. En grafos moleculares los nodos representan los átomos y las aristas representan vínculos químicos. Consiguientemente, los grafos moleculares expresan la estructura de compuestos orgánicos en términos de conectividad de átomos. Además, se pueden asociar estos grafos con diferentes clases de matrices numéricas en distintos niveles de organización estructural. La Matriz de adyacencia es quizás la más sencilla de explicitar. Estas matrices son tablas cuadradas (número de filas por número de columnas) de nxn elementos, donde n es el número de nodos del sistema. El elemento de una celda matricial es $b_{ij} = 1$ si el nodo $i^{th}$ está unido por una arista con el nodo $j^{th}$ en el grafo. Ello significa, que el átomo $i^{th}$ está químicamente enlazado al átomo $j^{th}$ en la estructura del compuesto químico en cuestión; por ejemplo, un fármaco.

Así, se pueden calcular fácilmente los diferentes parámetros numéricos de estas matrices que se asocian al grafo y que suelen describir la estructura de los fármacos. Cuando estos números están basados sólo en información de conectividad (patrones de enlace átomo-átomo) a menudo son nombrados como medidas de Conectividad o Índices Topológicos (TIs). De estas redes, se pueden calcular en primer lugar los TIs de átomos y, en este punto, se pueden combinar los TIs de los átomos para calcular los TIs de grupos atómicos o de todas las moléculas del compuesto químico. Posiblemente, el TI más sencillo para el átomo j-ésimo es el grado $\delta(j)$ del nodo correspondiente en el grafo (número de átomos enlazado a j = valencia del átomo). Una ventaja importante de los TIs es que la base teórica es sencilla de entender incluso para científicos no-expertos en técnicas computacionales. Una revisión excelente sobre modelos QSAR basados en TIs ha sido publicada por Garcia-Domenech, Galvez, de Julian-Ortiz, y Pogliani. También Todeschini y Consonni realizan una recopilación exhaustiva que sistematiza más de 1,600 descriptores moleculares para el descubrimiento de fármacos de molécula pequeña, incluyendo varios TIs. Es preciso aclarar que algunos de estos trabajos abordan temas muy similares. Muchos

investigadores definen TIs para grafos o redes que utilizan un procedimiento vector-matriz-vector (**vMv**). Ello indica semejanzas significativas entre estos sistemas. De hecho, el primer TI definido en un contexto químico, el índice Wiener W (ver ecuación 1), tiene esta forma. Además, muchos otros TIs pueden ser representados en la forma **vMv**; por ejemplo: los índices Zagreb M1 y M2, número H de Harary, índice de conectividad de Randič $\chi$, índice de Valencia o conectividad $\chi^v$, índice J o de Balaban, y los índices de auto-correlación de Broto–Moreau $ATS_d$.

$$W = \frac{1}{2}\left(\mathbf{u}\cdot\mathbf{D}\cdot\mathbf{u}^T\right) \quad (1) \qquad M_1 = \mathbf{v}\cdot\mathbf{A}\cdot\mathbf{u}^T \quad (2) \qquad M_2 = \frac{1}{2}\left(\mathbf{v}\cdot\mathbf{A}\cdot\mathbf{v}^T\right) \quad (3)$$

$$H = \frac{1}{2}\left(\mathbf{u}\cdot\mathbf{D}^{-k}\cdot\mathbf{u}^T\right) \quad (4) \qquad \chi = \mathbf{v'}\cdot\mathbf{A}\cdot\mathbf{v'}^T \quad (5) \qquad \chi^v = \mathbf{v''}\cdot\mathbf{A}\cdot\mathbf{v''}^T \quad (6)$$

$$J = \frac{1}{2}\cdot C\cdot\left(\mathbf{d'}\cdot\mathbf{A}\cdot\mathbf{d'}^T\right) \quad (7) \qquad ATS_d = \mathbf{w}\cdot^{\mathbf{m}}\mathbf{B}\cdot\mathbf{w}^T \quad (8)$$

Todos los vectores y las matrices utilizados en estas expresiones han sido exhaustivamente explicados en la literatura científica. La característica común de expresión **vMv** para la mayoría de los TIs, hace esperar que un gran número de investigadores ya no se decante por desarrollar nuevos índices de esta clase. En este sentido, no se sabe si se está adentrando en la era Fukuyama, el "fin de historia" para el desarrollo de nuevos TIs. De todas formas, queda claro que la aplicación de TIs es un campo de búsqueda creciente. Una discusión de este tema aparece en una reciente revisión punlicada por la doctoranda en colaboración con otros investigadores. Como quiera que la nota más importante de los TIs es, quizás, la posibilidad de buscar modelos QSAR/QSPR, no sólo para sistemas moleculares sino también para Redes Complejas *"Complex Networks"*, que representan sistemas en niveles estructurales más elevados como los biológicos, sociales, tecnológicos o legales.

En este contexto, un QSAR es una ecuación linear sencilla o un modelo complejo no-linear (basado, por ejemplo, en ANNs) para pronosticar la actividad farmacológica de una sustancia que utiliza TIs como entradas (inputs) para cuantificar la estructura química del fármaco representada por un grafo, ver **Figura 1**. Sea S, una variable real utilizada para cuantificar una actividad biológica dada, TI(t) un TI de tipo t

seleccionado entre c clases de TIs, y $a_t$ los coeficientes de estos TIs en el modelo QSAR, y $a_0$ el término independiente; se puede buscar una ecuación QSAR con la forma linear:

$$S = \sum_{t=1}^{c} a_t \cdot TI(t) + a_0 \quad (9)$$

El Análisis Discriminante Linear *"LDA",* es una de las técnicas más usadas en estudios QSAR con TIs. Los parámetros estadísticos para la ecuación LDA son: n = número de casos, U = Estadígrafo U o de λ de Wilks (U = 1 para discriminación perfecta y U = 0 para discriminación nula), F = Razón de Fisher, y p = el nivel de error. Se pueden utilizar diferentes software para calcular TIs. Uno de los programas más conocidos es la plataforma CODESSA, ver el trabajo de Karelson, Lobanov, y Katritzky. CODESSA PRO Por Katritzky, Karelson, y Petrukhin *et al.*; con la última actualización del software CODESSA clásico, ver enlace: http://www.codessa-pro.com/. Otro programa importante es el MODesLab, desarrollado por Estrada *et al.*, ver enlace: http://www.modeslab.com/?Casade página.php. Se puede mencionar también el software MOE [70], el cual incluye también módulos de cálculo QSAR y de TIs; ver enlace: http://www.chemcomp.com/software.htm. El uso de DRAGÓN o MOE para clacular TIs en QSAR ha sido revisado recientemente por Helguera *et al.* y Vilar *et al*. Es de justicia mencionar, además, el programa Molconn-Z desarrollado por Kier *et al.*, uno de los más útiles para el cálculo de TIs, ver enlace: http://www.edusoft-lc.com/molconn/. Molconn-Z proporciona 749 descriptores químicos y es el sucesor al programa Molconn-X, el cual generaba 462 descriptores. Recientemente, el grupo de Quimiometría de Milan (Milano Chemometrics) dirigido por Todeschini *et al.* ha liberado el servidor web EDRAGON. Este es una versión pública "online" de su famoso programa DRAGON. EDRAGON fué publicado por Tetko *et al.* y está disponible para uso público en: http://www.vcclab.org/lab/edragon/. Ambas versiones calculan 1600 parámetros moleculares que incluyen diferentes TIs. Por último, MARCH-INSIDE es un programa desarrollado por González-Díaz *et al*. (**Figura 1**), que utiliza la teoría de las cadenas de Markov para generar parámetros que

describen numéricamente la estructura química de los fármacos y sus dianas moleculares. Este será el programa usado para la mayoría de los estudios desarrollados en esta tesis.

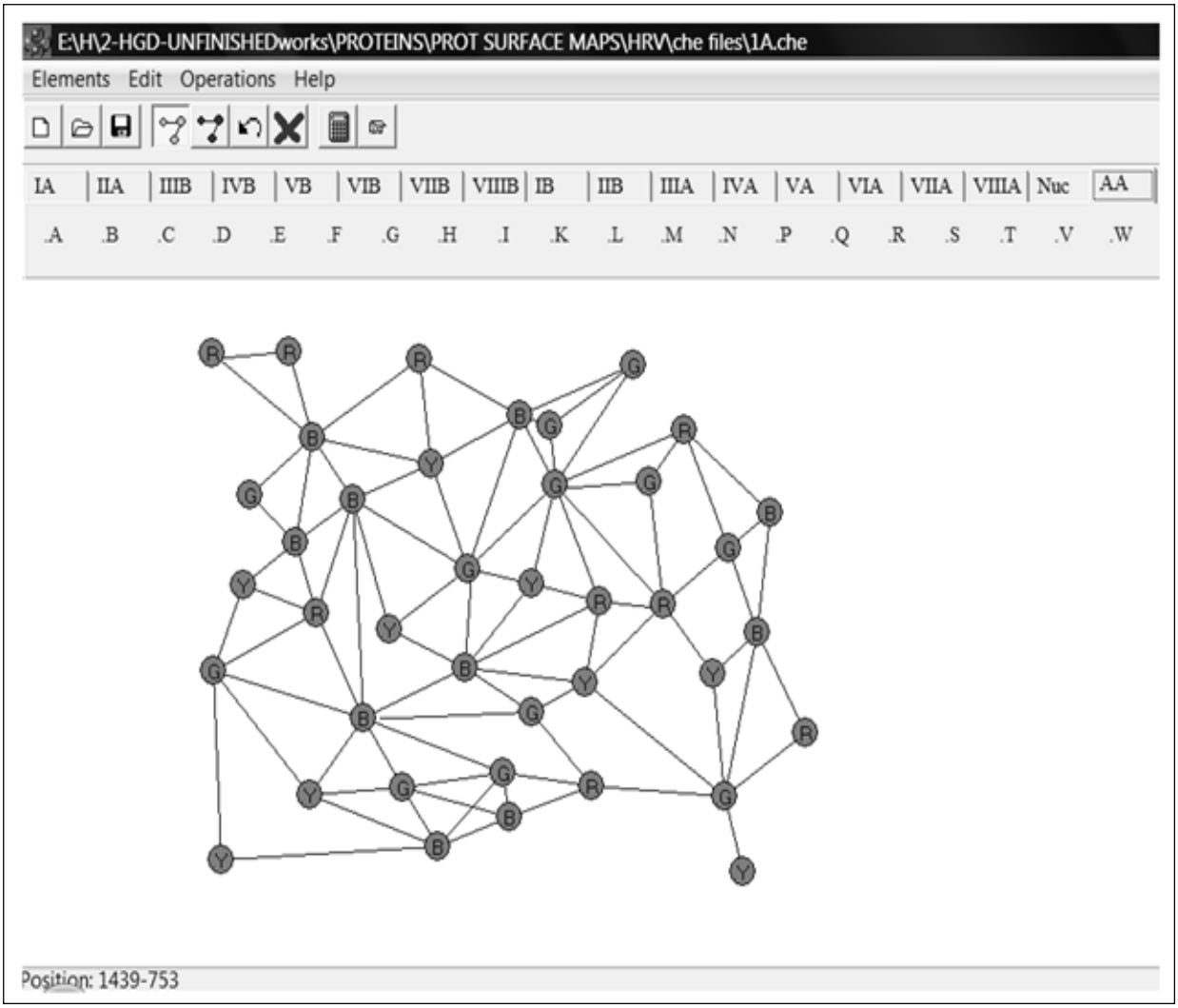



***Figura 1***. *Interfaz gráfica de la aplicación MARCH-INSIDE*

## Bibliografía Consultada

Altermann, E.; Klaenhammer, T.R. PathwayVoyager: pathway mapping using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. *BMC Genomics* **2005**; 6: 60.

Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; Harris, M.A.; Hill, D.P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J.C.; Richardson, J.E.; Ringwald,

M.; Rubin, G.M.; Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **2000**; 25: 25-9.

Devillers, J.; Balaban, A.T. Topological Indices and Related Descriptores in QSAR and QSPR. Gordon and Breach: The Netherlands **1999**.

Estrada, E. Generalization of topological indices. *Chem Phys Lett* **2001**; 336: 248-52.

Estrada, E.; Diaz, G.A.; Delgado, E.J. Predicting infinite dilution activity coefficients of organic compounds in water by quantum-connectivity descriptores. *J Comput Aided Mol Des* **2006**; 20: 539-48.

Estrada, E.; Uriarte, E. Quantitative structure--toxicity relationships using TOPS-MODE. 1. Nitrobenzene toxicity to Tetrahymena pyriformis. *SAR & QSAR in Environmental Research* **2001**; 12: 309-24.

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

Garcia-Domenech, R.; Galvez, J.; de Julian-Ortiz, J.V.; Pogliani, L. Some new trends in chemical grafo theory. *Chem Rev* **2008**; 108: 1127-69.

Gonzalez-Diaz, H.; Cruz-Monteagudo, M.; Vina, D.; Santana, L.; Uriarte, E.; De Clercq, E. QSAR for anti-RNA-virus activity, synthesis, and assay of anti-RSV carbonucleosides given a unified representation of spectral moments, quadratic, and topologic indices. *Bioorg Med Chem Lett* **2005**; 15: 1651-7.

González-Díaz, H.; Cruz-Monteagudo, M.; Vina, D.; Santana, L.; Uriarte, E.; De Clercq, E. QSAR for anti-RNA-virus activity, synthesis, and assay of anti-RSV carbonucleosides given a unified representation of spectral moments, quadratic, and topologic indices. *Bioorg Med Chem Lett* **2005**; 15: 1651-7.

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: 1025-39.

Hill, T.; Lewicki, P. STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining. StatSoft: Tulsa **2006**.

Hoffman, B.T.; Kopajtic, T.; Katz, J.L.; Newman, A.H. 2D QSAR modeling and preliminary database searching for dopamine transporter inhibitors using genetic algorithm variable selection of Molconn Z descriptores. *J Med Chem* **2000**; 43: 4151-9.

Jiao, S.; Bailey, C.P.; Zhang, S.; Ladunga, I. Probabilistic peak calling and controlling false discovery rate estimations in transcription factor binding site mapping from ChIP-seq. *Methods in Molecular Biology* **2010**; 674: 161-77.

Karelson, M.; Lobanov, V.S.; Katritzky, A.R. Quantum-Chemical Descriptores in QSAR/QSPR Studies. *Chem Rev* **1996**; 96: 1027-44.

Katritzky, A.R.; Kulshyn, O.V.; Stoyanova-Slavova, I.; Dobchev, D.A.; Kuanar, M.; Fara, D.C.; Karelson, M. Antimalarial activity: a QSAR modeling using CODESSA PRO software. *Bioorganic & Medicinal Chemistry* **2006**; 14: 2333-57.

Klovdahl, A.S.; Potterat, J.J.; Woodhouse, D.E.; Muth, J.B.; Muth, S.Q.; Darrow, W.W. Social networks and infectious disease: The Colorado Springs study. *Soc. Sci. Med.* **1994**; 38: 79-88.

Liljeros, F.; Edling, C.R.; Amaral, L.A.N.; Stanley, H.E.; Aberg, Y. The webof human sexual contacts. *Nature* **2001**; 411: 907-8.

Matamala, A.R.; Estrada, E. Generalised topological indices: Optimisation methodology and physico-chemical interpretation. *Chem Phys Lett* **2005**; 410 343–7.

Matamala, A.R.; Estrada, E. Simplex Optimization of Generalized Topological Index (GTI-Simplex): A Unified Approach to Optimize QSPR Models. *J Phys Chem A* **2005**; 109: 9890-5.

McKusick, V.A. Mendelian Inheritance in Man and its online version, OMIM. *American Journal of Human Genetics* **2007**; 80: 588-604.

Todeschini, R.; Consonni, V. Handbook of Molecular Descriptores. Wiley-VCH: **2002**.

## II.2. Redes de Neuronas Artificiales *(ANN "Artificial Neural Networks")*

En cualquier caso, la complejidad de las redes jurídicas y de los problemas derivados del SNA es muy elevada. Esto puede determinar que no siempre sea posible encontrar modelos QSPR lineares con suficiente poder predictivo. En este contexto, el uso de las ANN puede ser de gran utilidad. Las redes de neuronas artificiales (denominadas habitualmente como RNA o en idioma inglés como ANN) son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas que colaboran entre sí para producir un estímulo de salida. En inteligencia artificial es frecuente referirse a ellas como redes de neuronas artificiales, redes neuronales artificiales, sistemas neuromórficos o sistemas conexionistas. Las ANN pueden ser consideradas como redes formadas por nodos y aristas al igual que las redes complejas bio-moleculares, tecnológicas, y socio-jurídicas. Una de las diferencias es que las ANN pueden procesar y aprender a predecir información relacionada con otros fenómenos incluyendo fenómenos complejos como los representados por las redes bio-moleculares, tecnológicas y sociales-jurídicas. Una ANN se compone de unidades llamadas neuronas. Cada neurona recibe una serie de entradas a través de interconexiones y emite una salida. Esta salida viene dada por tres funciones (en la **Figura 2** muestra, de forma básica, el esquema general de una ANN):

1. Una función de propagación (también conocida como función de entrada), que por lo general consiste en una función sumatorio dónde cada entrada se multiplica por el peso de su interconexión o canal que la transmite. Si el peso es positivo, la conexión se denomina excitatoria; si es negativo, se denomina inhibitoria.
2. Una función de excitación o de activación, que toma como entrada la salida de la anterior. Puede no existir, siendo en este caso la salida la misma función de propagación. Las más comunes son las funciones tipo "umbral", las sigmoideas o las "hiperbólicas tangentes", en función del tipo de datos de entrada.

3. Una función de transferencia, o de salida, que se aplica al valor devuelto por la función de activación y es la salida que ofrece la ANN a sus entrada en cada momento. Se utiliza, por ejemplo, para acotar la salida de las neuronas y, generalmente, viene dada por la interpretación que se quiera dar a dichas salidas.

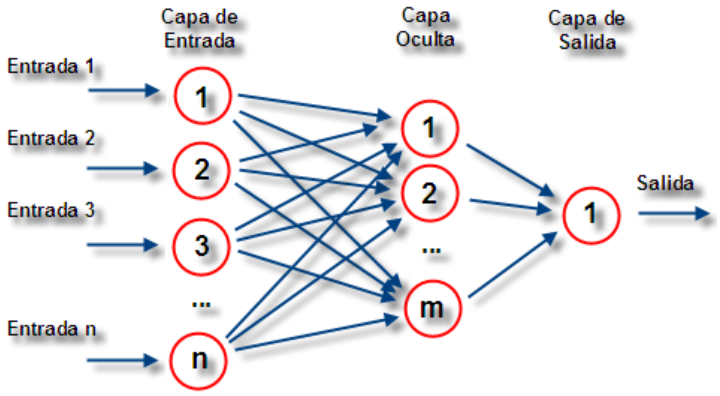



***Figura 2.*** *Esquema de la arquitectura de una ANN estándar.*

Por tanto, las ANNs pueden ser consideradas como redes capaces de procesar la información relacionada con otras redes complejas y predecirlas. El procedimiento para lograr este objetivo podría ser el mismo que el usado en un análisis QSPR si se sustituye el modelo linear por una ANN. La **Figura 3** muestra el esquema general de un estudio QSPR usando una ANN; aplicado a redes complejas bio-moleculares, sociales o jurídicas.

Las ANNs tienen muchas ventajas debidas, sobre todo, a que su arquitectura está basada en la estructura del sistema nervioso, principalmente el cerebro, ya la representación del conocimiento en las mismas es distribuida; esto es, reside en el peso de las conexiones entre los elementos, no en los propios nodos de la red, y tiene un comportamiento "holístico" por el hecho de que todos los elementos del sistema están conectados entre sí, no hay elementos o partes del sitema aisñados del resto. Además, tienen la capacidad de aprender a partir de ejemplos "socráticamente"; esto es, sin que nadie le proporcione instrucciones de lo que debe o no aprender.

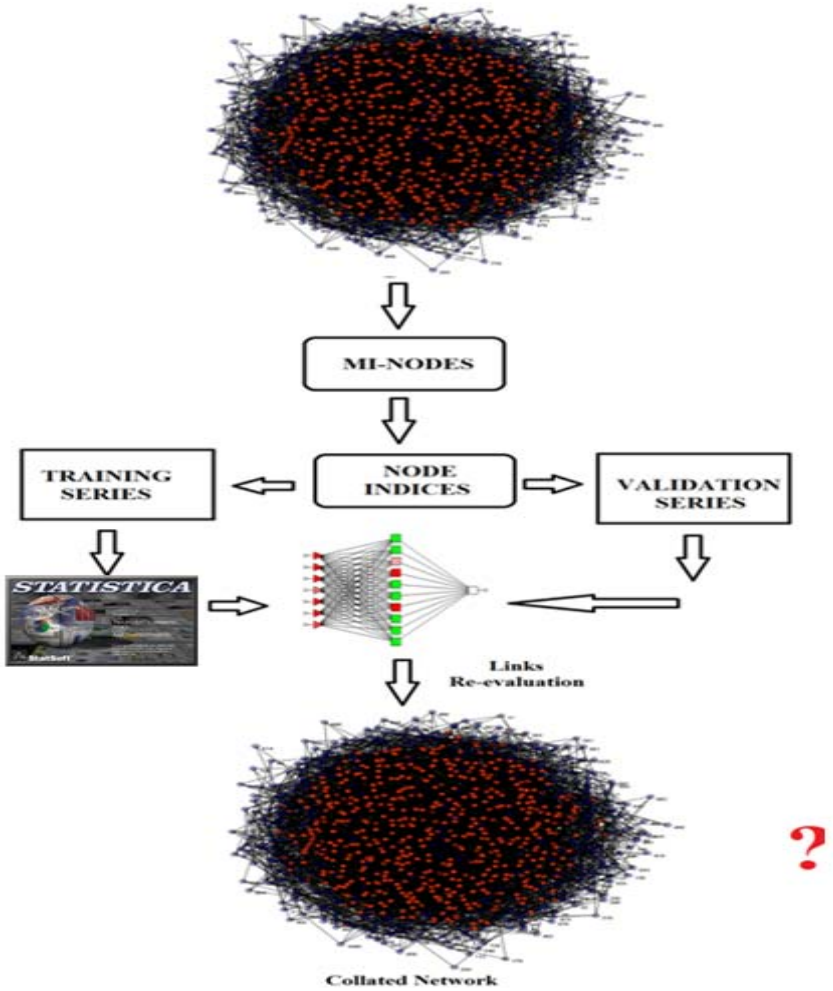



***Figura 3**. Esquema de un estudio ANN-QSPR de redes complejas bio-moleculares, sociales, o jurídicas.*

Estas ventajas las convierten en algoritmos muy poderosos para estudios QSPR. Algunas de dichas ventajas son:

1. Aprendizaje: mencionada con anterioridad. Las ANN pueden aprender mediante una etapa que se llama etapa de aprendizaje. Esta consiste en proporcionar a la ANN datos como entrada a su vez que se le indica cuál es la salida (respuesta) esperada. En el caso del QSPR los datos de entrada serían las centralidades de nodos de las redes complejas bio-molecular, tecnológica, o socio-jurídica que se quiere estudiar. La salida sería la magnitud S referida anteriormente en los modelos QSPR.

2. Auto organización: Una ANN crea su propia representación de la información en su interior, descargando al usuario de esto.
3. Tolerancia a fallos: Debido a que una RNA almacena la información de forma redundante, ésta puede seguir respondiendo de manera aceptable aun si se daña parcialmente.
4. Flexibilidad: Una ANN puede manejar cambios no importantes en la información de entrada, como señales con ruido u otros cambios en la entrada.
5. Tiempo real: La estructura de una ANN es paralela, por lo cual si esto es implementado con computadoras o en dispositivos electrónicos especiales, se pueden obtener respuestas en tiempo real.

En un reciente trabajo Duardo-Sanchez *et al*., aplicaron diferentes tipos de ANN a la predicción de la causalidad penal. Los resultados encontrados fueron prometedores, ver **Tabla 1**.

***Tabla 1.*** *Resultados de la ANN*

| Clasificadores | | Rendimiento | | Neuronas | | |
|---|---|---|---|---|---|---|
| ANN | Perfil | Entrenamiento | Validación | $E_0$ | $H_1$ | $H_2$ |
| 1 | MLP 6:6-11-11-3:1 | 87.8 | 76.9 | 6 | 11 | 11 |
| 2 | PNN 6:6-41-3:1 | 70.7 | 69.2 | 6 | 41 | 0 |
| 3 | LNN 5:5-3:1 | 87.8 | 76.9 | 5 | 0 | 0 |
| 4 | RBF 5:5-6-3:1 | 85.4 | 76.9 | 5 | 6 | 0 |
| 5 | LNN 6:6-3:1 | 87.8 | 76.9 | 6 | 0 | 0 |

*Un número de neuronas en la Entrada ($E_0$), Ocultas capa 1 (H1), Ocultas capa 2 (H2).*

Los modelos mejores pueden discriminar correctamente el 87% de las causas de delito entre otras causas de delito potencial en 17 casos de delito. En la **Figura 4**, ilustramos la topología de algunos de la ANN modelos entrenaron en este estudio.

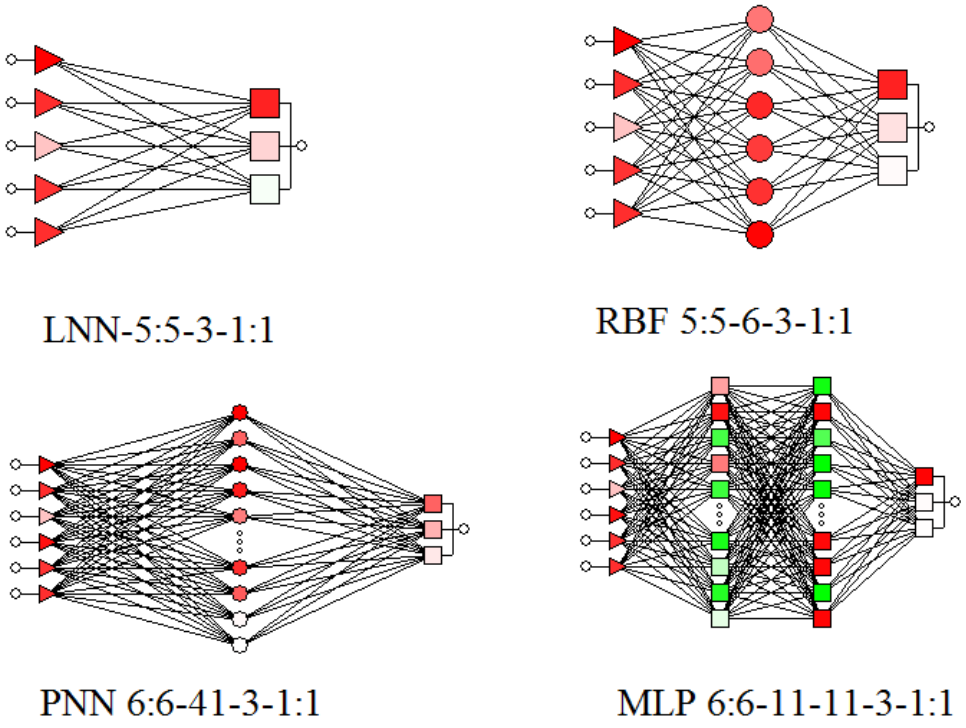



***Figura 4.*** *Topología de algunas de las ANNs probadas en este trabajo*

**Bibliografía**

Duardo-Sanchez, A. Criminal law networks, markov chains, Shannon entropy and artificial neural networks. In *Complex Network Entropy: From Molecules to Biology, Parasitology, Technology, Social, Legal, and Neurosciences*, González-Díaz, H.; Prado-Prado, F.J.; García-Mera, X., Eds. Transworld Research Network: Kerala, India, **2011**; pp 107-14.

Duardo-Sanchez, A. Study of criminal law networks with Markov-probability centralities. In *Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks*, González-Díaz, H.; Munteanu, C.R., Eds. Transworld Research Network: Kerala, India, **2010**; pp 205-12.

Riera-Fernandez, P.; Munteanu, C. R.; Escobar, M.; Prado-Prado, F.; Martin-Romalde, R.; Pereira, D.; Villalba, K.; Duardo-Sanchez, A.; Gonzalez-Diaz, H., New Markov-Shannon Entropy models to assess connectivity quality in complex

networks: from molecular to cellular pathway, Parasite-Host, Neural, Industry, and Legal-Social networks. *J Theor Biol* **2012,** *293*, 174-88.

Riera-Fernandez, P.; Munteanu, C. R.; Pedreira-Souto, N.; Martin-Romalde, R.; Duardo-Sanchez, A.; Gonzalez-Diaz, 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.

## II.3. Representaciones generales de redes complejas

Un grafo $G=(V, A)$ consiste en dos conjuntos $V$ y $A$ tales que $V \neq \emptyset$ y $A$ es un conjunto de pares de elementos de $V$. Los elementos de $V \equiv \{v_1,v_2...v_N\}$ son los vértices o nodos del grafo $G$, mientras que los elementos de $A \equiv \{a_1,a_2...a_K\}$ son sus aristas o conexiones. El número de elementos de $V$ y $A$ están denotados por $N$ y $K$ respectivamente (**Figura 2**). Se dice que un grafo es bipartito si V está compuesto por dos subconjuntos $V_A$ y $V_B$ tal que cada arista conecta un nodo de $V_A$ con un nodo de $V_B$. En algunos casos, es necesario asignar un sentido a las aristas; por ejemplo, si se quiere representar la red de las calles de una ciudad con sus direcciones únicas. En este caso se habla de un grafo dirigido y se usa el término arco para referirse a las aristas. En un grafo dirigido $a_{ij} \neq a_{ji}$ dos conceptos a tener en cuenta son los de: adyacencia e incidencia.

Adyacencia es la relación entre dos vértices. En el caso de un grafo no dirigido, si dos de ellos están conectados se dice que son adyacentes (la relación de adyacencia es simétrica), mientras que en el caso de un grafo dirigido, si $v_1$ se conecta con $v_2$ , se dice que $v_2$ es adyacente a $v_1$ . Así, si en un grafo todos los vértices son adyacentes entre si se dice que es completo.

Incidencia se refiere a la relación que se establece entre un vértice y una arista. En un grafo dirigido de $v_1$ a $v_2$ se dice que el arco es incidente positivo con respecto a $v_1$ (sale de $v_1$) y que es incidente negativo respecto a $v_2$ (llega a $v_2$). Al número de aristas incidentes sobre un vértice se le conoce como grado y, en el caso de los grafos dirigidos, se divide en dos tipos: el grado positivo, que es el número de arcos que parten del vértice, y el grado negativo, que es el número de arcos que llegan al vértice. La distancia entre dos nodos se define como el número mínimo de aristas que los separa y la excentricidad como la distancia máxima que se puede recorrer en $G$ partiendo de un vértice determinado. El diámetro de $G$ se corresponde con la máxima excentricidad y el radio con la mínima. La densidad se define como el número de aristas del grafo dividido entre el número total de aristas posibles.

Generalmente, existen dos formas de representar la información de los grafos (ver **Figura 5**). La primera de ellas es el uso de matrices, que pueden ser:

a) Matrices de incidencia: El grafo ($G$) está representado por una matriz de $N$ vértices por $K$ aristas, donde el par i-j (arista-vértice) contiene la información de la arista (1 - conectado, 0 - no conectado).
b) Matrices de adyacencia: $G$ está representado por una matriz cuadrada de tamaño $N^2$, donde $N$ es el número de vértices. Si hay una arista entre un vértice i y un vértice j, entonces el elemento $a_{i,j}$ es 1, de lo contrario, es 0.

La segunda es el uso de listas, que pueden ser:

a) Listas de incidencia: Se utiliza una lista de pares de vértices (ordenados si el grafo es dirigido), donde cada par representa una de las aristas.
b) Listas de adyacencia: Cada vértice tiene una lista de vértices que son adyacentes a él. Esto causa redundancia en un grafo no dirigido (ya que si $v_1$ y $v_2$ están conectados $v_1$ existe en la lista de adyacencia de $v_2$ y viceversa).

Desde el punto de vista del manejo de los datos, las listas son preferidas en grafos dispersos porque tienen un uso de la memoria más eficiente. Las matrices permiten un acceso rápido a los datos, pero pueden consumir grandes cantidades de memoria. En muchos casos, los términos grafo y red se usan indistintamente, aunque no significan exactamente lo mismo. Cuando se habla de grafo se refiere a un objeto matemático, mientras que cuando se usa el término red se refiere a un sistema real (como, por ejemplo, una red de ordenadores o una red social) en el que los vértices se corresponden con las entidades reales que se pretenden representar (ordenadores o personas) y las aristas con las relaciones de distinta naturaleza que se establecen entre ellos (conexiones por cable, relaciones de amistad, etc.).

Durante las dos últimas décadas, el avance en el conocimiento de distintos sistemas complejos, la creación de bases de datos disponibles para la comunidad científica y el desarrollo de herramientas informáticas capaces de manejar de forma eficiente grandes cantidades de datos ha permitido caracterizar de forma eficiente la estructura de redes complejas de sistemas muy diversos. Así, se ha observado que las características de

muchas redes reales son diferentes de las que presentan las redes totalmente regulares (en las que todos sus vértices presentan el mismo grado) o totalmente aleatorias usadas normalmente como modelos en el ámbito de la teoría de grafos.

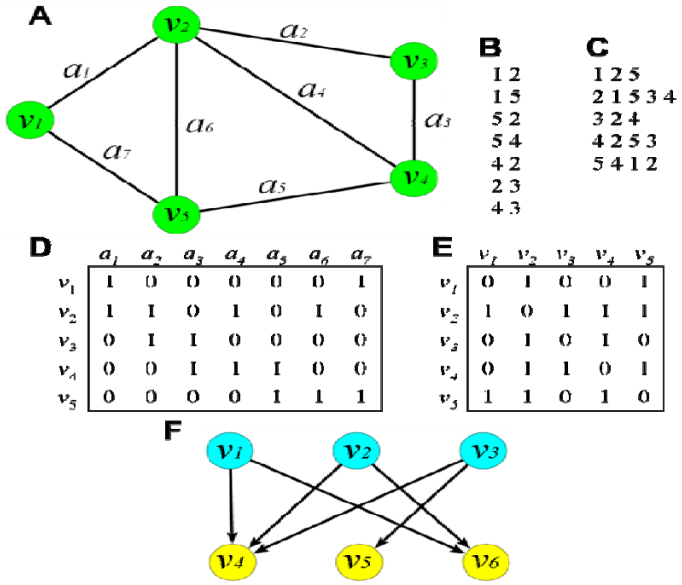



***Figura 5.*** *A: Grafo no dirigido. v: vertices, a: aristas. B: Lista de incidencia. C: Lista de adyacencia. D: Matriz de incidencia. E: Matriz de adyacencia. F: Grafo bipartito y dirigido.*

En concreto, el estudio de varias redes reales de naturaleza muy distinta y de sus propiedades ha llevado a proponer dos tipos de redes:

a) Redes de mundo pequeño (*Small world networks*):

Se trata de de redes que presentan un coeficiente de agrupamiento elevado (este coeficiente cuantifica la interconexión o agrupamiento de un nodo con sus vecinos) y una distancia media entre nodos relativamente pequeña (**Figura 3**). Las primeras redes de mundo pequeño fueron empleadas en experimentos de sociología. De forma especial, se pueden destacar los experimentos llevados a cabo por Milgram en la década de los 60, en los que se intentaba averiguar el número de conocidos que separan a dos personas determinadas. Estos experimentos dieron lugar a los conceptos de "mundo pequeño" y de los "seis grados de separación" (dos personas cualesquiera están separadas por una cadena de conocidos de distancia media 6). Posteriormente Watts y Strogatz, estudiaron

la red neural del nemátodo *Caenorhabditis elegans*, la red de colaboraciones de actores en películas y la red eléctrica del oeste de EE.UU., demostrando que estas redes presentan una estructura de mundo pequeño (es en este estudio donde se acuña el término "red de mundo pequeño") y que dicha estructura se encuentra entre los extremos de lo totalmente regular y lo totalmente aleatorio. A partir de este estudio, se llevaron a cabo una gran cantidad de trabajos encaminados a comprender este tipo de redes y sus propiedades.

b) Redes libres de escala (*Scale free networks*):

Las redes libres de escala se caracterizan porque en ellas algunos vértices están altamente conectados (a estos nodos con grado alto se les denomina *hubs*), aunque en general el grado de conexión de casi todos los vértices de la red es bastante bajo (**Figura 6**). En general, se puede decir que la probabilidad de que un vértice de la red esté conectado con $k$ vértices $P(k)$ es proporcional a $k^{-\gamma}$, es decir sigue una ley de potencias. El exponente $\gamma$ es específico de la red estudiada y, generalmente, su valor se encuentra en el rango $2 < \gamma \leq 3$. El interés por el estudio de redes con estas características comenzó con los trabajos de Barabasi *et al*. acerca de la topología de internet y ha continuado con el estudio de otras redes complejas pertenecientes a distintos campos, como por ejemplo la red de colaboraciones de actores, las redes de aerolineas de estados unidos, las redes metabólicas, las redes de relaciones sexuales, etc. Es importante aclarar que una red puede ser de mundo pequeño y libre de escala a la vez.

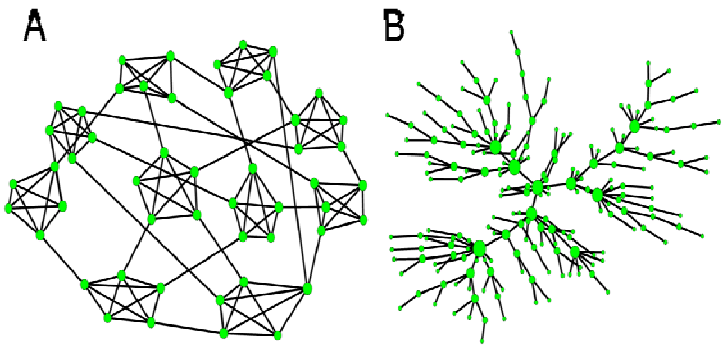



***Figura 6.*** *Ejemplos de redes complejas. A: Red de mundo pequeño. B: Red libre de escala.*

**Bibliografía Consultada**


Amaral, L. A. N.; Ottino, J. M., Complex networks: Augmenting the framework for the study of complex systems. *Eur. Phys. J. B* **2004,** *38*, 147–162.

Barabasi, A. L.; Albert, R., Emergence of scaling in random networks. *Science* **1999,** *286* (5439), 509-12.

Barabasi, A. L.; Bonabeau, E., Scale-free networks. *Sci. Am.* **2003,** *288* (5), 50-59.

Barthélémy, M.; Nunes Amaral, L. A., Small-World Networks: Evidence for a Crossover Picture. *Phys. Rev. Lett.* **1999,** *82* (15), 3180–3183; Barrat, A.; Weigt, M., On the properties of small-world network models. *Eur. Phys. J.* **2000,** *13*, 547{560; Amaral, L. A.; Scala, A.; Barthelemy, M.; Stanley, H. E., Classes of small-world networks. *Proc. Natl. Acad. Sci. U. S. A.* **2000,** *97* (21), 11149-52.

Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; Hwang, D. U., Complex networks: Structure and dynamics. *Phys. Rep.* **2006,** *424*, 175-308.

Bornholdt, S.; Schuster, H. G., *Handbook of Graphs and Complex Networks: From the Genome to the Internet*. WILEY-VCH GmbH & CO. KGa.: Wheinheim, 2003; Newman, M., The Structure and Function of Complex Networks. *SIAM Review* **2003,** (56), 167-256.

Cormen, T. H.; Leiserson, C. E.; Rivest, R. L.; Stein, C., *Introduction to Algorithms*. 3th ed.; The MIT Press: Cambridge Massachusetts, **2009**; p 1312.

Dehmer, M., *Structural analysis of complex networks*. 1st ed.; Birkhäuser: Boston, **2010**; p 492.

Dehmer, M.; Emmert-Streib, F., *Analysis of complex networks: From biology to linguistics*. Wiley-Blackwell: Wheinheim, **2009**; p 462.

Diestel, R., *Graph Theory*. 2nd ed.; Springer-Verlag: New York, 2000.

Euler, L., Solutio problematis ad geometriam situs pertinentis. *Comentarii Academiae Scientiarum Imperialis Petropolitanae* **1736,** *8*, 128-140.

Junker, B. H.; Schreiber, F., *Analysis of biological networks*. Wiley & Sons: New Jersey, 2008; p 346; Estrada, E.; Fox, M.; Higham, D. J.; Oppo, G. L., *Network Science: Complexity in Nature and Technology*. Springer: London, **2010**; p 256.

König, D., *Theorie der endlichen und unendlichen Graphen: Kombinatorische Topologie der Streckenkomplexe* Akademische Verlags gesellschaft Leipzig, **1936**.

Loudon, K., *Mastering Algorithms with C*. 1st ed.; O'Reilly Media: Sebastopol, California, **1999**; p 560.

Milgram, S., The small world problem. *Psychology today* **1967,** *1* (1), 61-67.

Réka, A.; Barabasi, A.L., Statistical mechanics of complex networks. *Rev Mod Phys* **2002,** *74* (1), 47-97.

Wasserman, S.; Faust, K., *Social network analysis: Methods and applications*. Cambridge University Press: Cambridge, **1999**.

Watts, D. J.; Strogatz, S. H., Collective dynamics of 'small-world' networks. *Nature* **1998,** *393*, 440-442.

## II.4. Representación de redes complejas en Ciencias Jurídicas

Un ejemplo muy interesante de la aplicación de las redes complejas al campo del Derecho se encuentra en los trabajos desarrollados por Fowler y sus colaboradores, quienes han creado una medida cuantitiva de la relevancia de una ley basada en patrones de cita de precedentes de Tribunal Supremo de EE.UU. Para este propósito, los autores consideraron las opiniones (casos) del Tribunal Supremo como nodos en una red jurídico-legal. Estos nodos de caso están enlazados a otros nodos de caso a través de la existencia de una cita como precedente. Los enlaces entre casos pueden tomar una de dos formas: "*a priori"* o "*a posteriori"*.

Las conexiones "*a priori"* son todos los precedentes que el caso cita *"outwards citation"*, mientras todas las opiniones subsiguientes que citan el caso son conexiones "*a posteriori"* o *"inwards citation"*. La combinación de nodos y enlaces ("*a priori"* y "*a posteriori"*) crea una red de precedente de cualquier número de casos. Por ejemplo, dentro de un área de asunto, entre todos los casos existentes. Ellos demostraron que un SNA de precedentes del Tribunal Supremo es una manera ideal de medir la importancia de un caso, viéndose como un concepto que se posiciona en el centro de cómo se perciben las decisiones de los jueces.

A modo ilustrativo, se puede considerar una red de precedentes de cinco decisiones la **Figura 7**, cada decisión es un nodo, mientras las flechas representan precedentes citados que van desde caso en cuestión hasta el caso citado. Notese que, cada caso, hace referencia a *Roe vs. Wade* (1973) pero que *Roe* no cita a otros casos mostrados. Esto significa que *Roe* tiene 4 citas *"a posteriori"* y 0 citas "*a priori"*. En contraste, *Planned Parenthood of Southeastern PA v. Casey* (1992) cita todos los otros casos mostrados, pero no es citado por ellos ya que fue el último de los cinco casos decidido. Así, *Casey* tiene 0 citas "*a posteriori"* y 4 "*a priori"*. Utilizando esta técnica, los autores buscan crear la red de todas las opiniones del Tribunal Supremo de los EE.UU. entre 1792 y 2005.

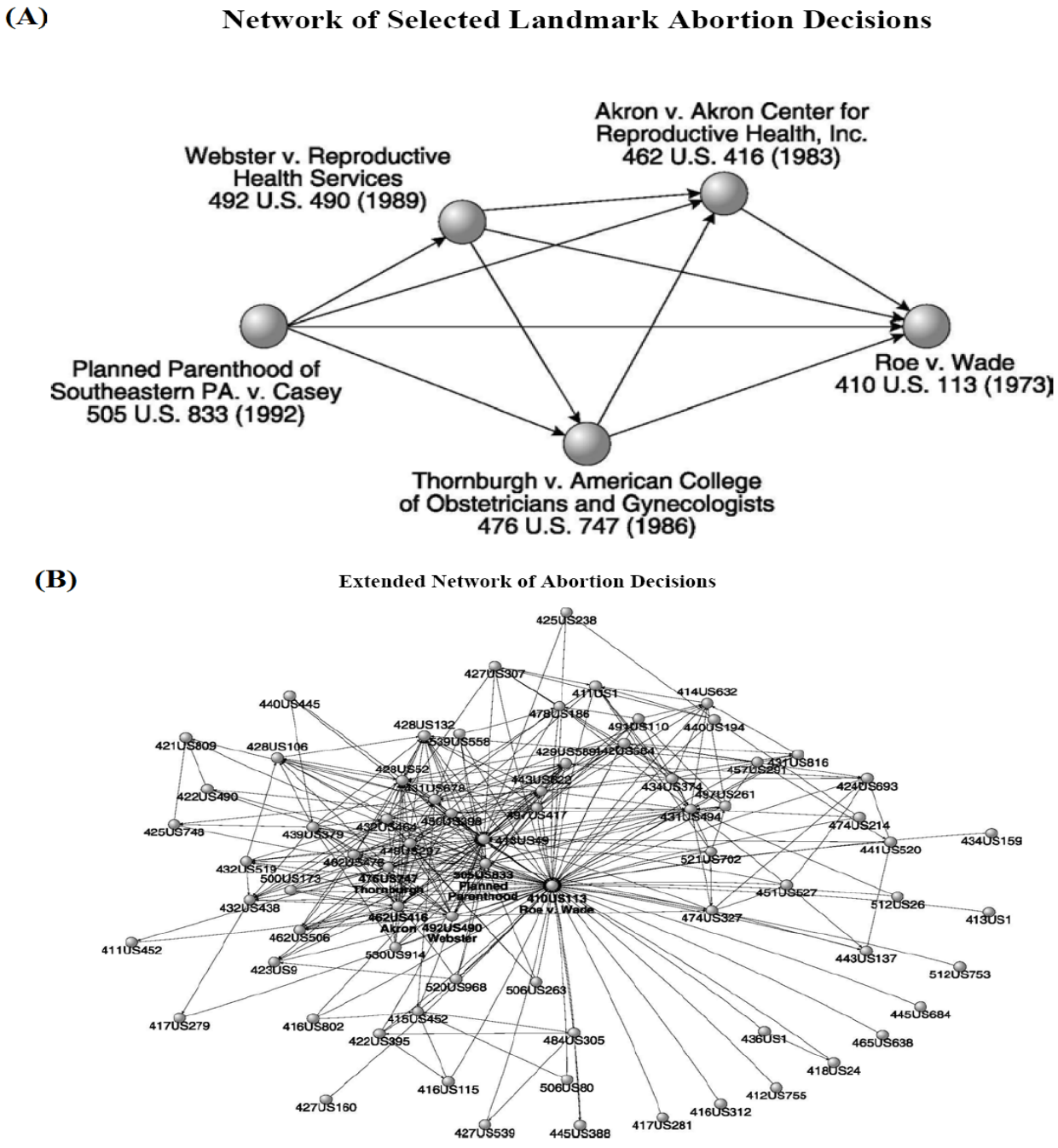



***Figura 7.*** *Ejemplos de redes en Ciencias Jurídicas.*

## Bibliografía

Fowler, J.H.; Jeon, S. The authority of Supreme Court precedent. *Social Networks* **2008**; 30: 16-30.

## II.5. Centralidades de nodos para redes jurídicas

En principio, es posible utilizar diferentes TIs para estudiar redes jurídicas, así como con cualquier otra red molecular, social, o red tecnológica. Se piensa que, en el caso particular de las redes jurídicas, así como en otras redes sociales abordadas en SNA, las centralidades de nodo pueden jugar un papel muy importante. La centralidad de un nodo en teoría de grafos y redes complejas se refiere al parámetro numérico que mide de algún modo la importancia relativa del nodo dentro de la red.

En SNA, el valor de centralidad de un nodo es útil para detectar las personas con mayor relevancia en una red social. En el contexto de esta tesis, las centralidades de nodo se consideran un caso particular de TIs. Como se mencionó anteriormente, las centralidades de nodo son equivalentes a un TI calculado sólo para un nodo y no para todo el grafo o parte de este. Bavelas introdujo el concepto de centralidad en 1940. Es uno de los conceptos más usados en SNA de los años 70; de ahí que muchos de los términos empleados en la definición de medidas de centralidad reflejan su origen sociológico. Este hecho es análogo a lo ocurrido con los TIs globales de un grafo; los cuales tuvieron un origen en la química matemática.

Contrariamente a la autoestima, la temperatura, el ingreso monetario, *etc.*, los valores de centralidad de nodo no constituyen unos atributos intrínsecos de los individuos o actores la red compleja. Por el contrario, los valores de centralidad de nodo son un atributo estructural; es decir, dependen estrictamente de las conexiones del nodo (ubicación en la red). En este sentido, para definir una centralidad se debe encontrar un parámetro que cuantifique la contribución de un nodo a la red según su ubicación en esta. Por ejemplo, en un grafo con topología tipo estrella el nodo central debería ocupar un valor máximo de centralidad, mientras que los nodos colgantes tendrían un valor de centralidad inferior.

En la **Tabla 2** se resumen algunos de los TIs más comúnmente utilizados para describir diferentes redes, desde moleculares a sociales, incluyendo por supesto las redes jurídicas.

***Tabla 2**. Centralidades de nodos utilizadas para describir redes complejas*

| Nombre | Formula | Software |
|---|---|---|
| Degree | $C_{\deg}(j)=\deg(j)$ | CBI |
| Eccentricity | $C_{ecc}(j)=\max\{dist(i,j)\}^{-1}$ | CBI |
| Closeness | $C_{clo}(j)=\left(\sum_{j\in V} dist(i,j)\right)^{-1}$ | CBI |
| Radiality | $C_{rad}(j)=\sum_{w\in V}(\Delta_G+1-dist(i,j))/(n-1)$ | CBI |
| Centroid Values | $C_{cen}(j)=\min\{f(i,j): i\in V\setminus\{\mathrm{j}\}\}$ | CBI |
| Stress | $C_{str}(j)=\sum_{s\notin v\in V}\sum_{t\notin v\in V}\sigma_{st}(j)$ | CBI |
| Shortest-path Betweenness | $C_{spb}(j)=\sum_{s\notin j\in V}\sum_{t\notin j\in V}\delta_{st}(j)$ | CBI |
| Current-Flow Closeness | $C_{cfc}(j)=(n-1)/\left(\sum_{i\notin V} p_{ji}(j)-p_{ij}(i)\right)$ | CBI |
| Current-Flow Betweenness | $C_{cfb}(j)=\sum_{s,t\in V}\tau_{st}(j)/(n-1)(n-2)$ | CBI |
| Katz Status Index | $C_{katz}=\sum_{k=1}^{\infty}\alpha^{k}\cdot(\mathbf{A}^{t})^{k}\cdot\mathbf{u}$ | CBI |
| Eigenvector | $EC(j)=e_1(j)$ | CBI |
| Closeness Vitality | $C_{clv}(j)=W(G)-W(G\setminus\{\mathrm{j}\})$ | CBI |
| Markov-Randic | ${}^{k}C_{\chi}(j)=\sum_{i}^{\delta_j}(\delta_i\cdot\delta_j)^{1/2}\cdot{}^{k}p_{ij}$ | MI |
| Markov-Shannon entropy | ${}^{k}C_{\theta}(j)=-\sum_{i}^{n}({}^{k}p_j)\cdot\log({}^{k}p_j)$ | MI |
| Markov Spectral moments | ${}^{k}C_{\pi}(j)=\sum_{i=j}^{n}{}^{k}p_{ij}=Tr\left[({}^{1}\Pi)^{k}\right]$ | MI |
| Markov-Harary | ${}^{k}C_{H}(j)=\frac{1}{2}\sum_{i}^{\delta_j}{}^{k}p_{ij}^{-1}$ | MI |
| Markov-Galvez | ${}^{k}C_{G}(j)=\frac{1}{2}\sum_{i,j}^{n}\left|{}^{k}CT_{ij}\right|\cdot\delta_j$ | MI |
| Markov-Rucker | ${}^{k}C_{WC}(j)=\frac{1}{2}\sum_{i}^{\delta_j}{}^{k}p_{ij}$ | MI |
| Markov-BM Autocorrelation | ${}^{k}C_{BM}(j)=\frac{1}{2}\sum_{i}^{\delta_j}{}^{k}p_{ij}\cdot{}^{k}p_{ji}$ | MI |
| Markov-Wiener | ${}^{k}C_{w}(j)=\frac{1}{2}\cdot\sum_{i\to 1}^{\delta_j}{}^{k}p_{ij}\cdot\mathrm{d}_{ij}$ | MI |
| Markov-Balaban | ${}^{k}C_{J}(j)=\frac{q}{\mu+1}\cdot\sum_{i\to j}^{\delta j}({}^{k}p_{ij}\cdot S_i\cdot S_j)^{-1/2}$ | MI |

[a] *G = (V, E) es un grafo no dirigido o dirigido con n = |V| vértices; δ(v) denota el grado del vértice o nodo v en un grafo no dirigido; dist(v, w) denota la distancia (camino más corto) entre los vertices v y w.* ***D*** *y* ***A*** *son las matrices de distancia topológica y adyacencia entre nodos del grafo G. Software: CB = CentiBin, PJK = Pajek, MI = MI-NODES.*

Además, en la **Figura 8**, se muestra, en un código de colores, los valores de una centralidad de nodo conocida como intermediación *"betweenness centrality"*.

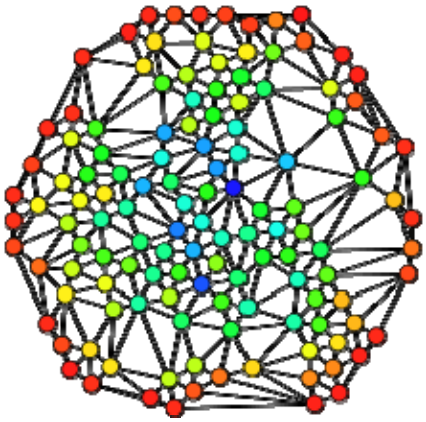


***Figura 8**. Valores de intermediación de los nodos en un grafo. Las tonalidades que van desde el rojo (valor 0) hasta el azul (valor máximo) indican la intermediación de los nodos en el grafo.*

**Bibliografía**


Bavelas, A. A mathematical model for group structures. Human Organization **1948**, 7: pp. 16-30.

Bonacich, P.. Factoring and weighting approaches to clique identification. Journal of Mathematical Sociology, **1972**, 2 (1): pp. 113-120.

Bonacich, P. (1987). Power and centrality: a family of measures. American Journal of Sociology 92 (5): pp. 1170-1182.

Bonacich, Phillip; Lloyd, Paulette. Eigenvector-like measures of centrality for asymmetric relations. Social Networks **2001**, 23 (3): 191-201.

Borgatti, S.P.; Everett, M.G. A graph-theoretic perspective on centrality. Social networks **2006**, 28 (4): pp. 466-484.

Borgatti, Stephen P. Centrality and network flow. Social Networks **2005**, 27: pp. 55-71.

Brandes, U. A faster algorithm for betweenness centrality. Journal of Mathematical Sociology **2001**, 25: pp. 163-177.

Dangalchev, Ch. Residual closeness in networks. Physica A **2006**, 365 (2): pp. 556-564.

David Austin. How Google Finds Your Needle in the Web's Haystack (en inglés).

Estrada, E.; Rodriguez-Velazquez, J.A. Subgrafo centrality in complex networks. Phys Rev E **2005**; 71: 056103.

Freeman, L. A set of measures of centrality based upon betweenness. Sociometry **1977**, 40 (1): pp. 35-41.

Freeman, L.C. Centrality in networks: I. Conceptual clarification. Social Networks **1979**, 1: pp. 215-239.

Freeman, L.C.; Borgatti, S.P.; White, D.R. Centrality in valued graphs: a measure of betweenness based on network flow. Social Networks **1991**, 13: pp. 141-154.

Gonzalez-Diaz, H.; Riera-Fernandez, P., New Markov-autocorrelation indices for re-evaluation of links in chemical and biological complex networks used in metabolomics, parasitology, neurosciences, and epidemiology. J Chem Inf Model **2012,** 52 (12), 3331-40.

González-Díaz, H.; Riera-Fernández, P.; Pazos, A.; Munteanu, C. R., The Rücker-Markov invariants of complex Bio-Systems: Applications in Parasitology and Neuroinformatics. Biosystems **2013,** 111 (3), 199-207.

Hubbell, C. (1965). An input-output approach to clique identification. Sociometry 28 (4): pp. 377-399. Beauchamp, M. A. An improved index of centrality. Systems Research and Behavioral Science **1965**, 10 (2): pp. 161-163.

Jimeng, Sun; Jie, Tang Charu C. Aggarwal. ed. A survey of models and algorithms for social influence analysis. Social network data analytics (Nueva York: Springer): **2011**, pp. 177-214.

Junker, B.H.; Koschuetzki, D.; Schreiber, F. Exploration of biological network centralities with CentiBiN. BMC Bioinformatics **2006**; 7: 219.

Katz, L. (1953). A new status index derived from sociometric index. Psychometrika 18 (1): pp. 39-43.

Newman, M.E.J. (2005). A measure of betweenness centrality based on random walks. Social Networks 27: pp. 39-54.

Newman, M.E.J. (2010). Networks: An Introduction. Oxford, Reino Unido: Oxford University Press.

Opsahl, T.; Agneessens, F.; Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks 32 (3): pp. 245.

Riera-Fernandez, P.; Munteanu, C. R.; Dorado, J.; Martin-Romalde, R.; Duardo-Sanchez, A.; Gonzalez-Diaz, H., From Chemical Graphs in Computer-Aided Drug Design to General Markov-Galvez Indices of Drug-Target, Proteome, Drug-Parasitic Disease, Technological, and Social-Legal Networks. Current Computer-Aided Drug Design **2011,** 7 (4), 315-337.

Riera-Fernandez, P.; Munteanu, C. R.; Escobar, M.; Prado-Prado, F.; Martin-Romalde, R.; Pereira, D.; Villalba, K.; Duardo-Sanchez, A.; Gonzalez-Diaz, H., New Markov-Shannon Entropy models to assess connectivity quality in complex networks: from molecular to cellular pathway, Parasite-Host, Neural, Industry, and Legal-Social networks. J Theor Biol **2012,** 293, 174-88.

Riera-Fernandez, P.; Munteanu, C. R.; Pedreira-Souto, N.; Martin-Romalde, R.; Duardo-Sanchez, A.; Gonzalez-Diaz, 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.

Sabidussi, G. The centrality index of a graph. Psychometrika **1966**, 31 (4): pp. 581-603.

Stephenson, K. A.; Zelen, M. Rethinking centrality: Methods and examples. Social Networks **1989**, 11 (1): pp. 1-37.

## II.6. Modelos QSPR de redes jurídicas

En el contexto de esta tesis y debido a su gran relevancia social, la predicción de la causalidad delictiva a partir del conocimiento que se tiene sobre los hechos acaecidos es de la mayor importancia. Diferentes medidas de causalidad de delito han sido desarrolladas antes por Devah en 2003. En un trabajo reciente, se ha utilizado la centralidad ${}^{k}\theta(j)$ basada en la entropía de una cadena de Markov representada por una matriz estocástica. Dicha matriz estocástica está asociada a una representación grafo-teórica del delito donde cada imputado y otros sujetos relacionados con este son representados por un nodo. Al mismo tiempo, se proponen como nuevas medidas de causalidad delictiva la suma de todos los valores de ${}^{k}\theta(j)$ del mismo orden k para todos los nodos colocados en el camino más corto que conecta el nodo original $n_i$ (causa posible) con el nodo final $n_{ii}$ (consecuencia). Así, se muestra como una de las aplicaciones más prometedoras el uso de este tipo de centralidad para encontrar modelos QSPR, capaces de detectar la causa principal de un delito que tenga una representación en forma de grafo con múltiples concausas.

En la **Figura 9,** se muestra un diagrama de flujo general para los plazos principales dados en un estudio QSPR, aplicado en SNA incluyendo redes jurídicas y otras. En otros trabajos recientes se han desarrollado representaciones de redes complejas similares a las ya discutidas en las secciones anteriores. En estas otras redes complejas jurídicas, los nodos de la red son leyes del sistema tributario español aprobadas en una fecha determinada. Una de las redes posibles a construir es aquella que mide la evolución en el tiempo de dicho sistema: dinámica y estabilidad del sistema de leyes tributarias. Para estos, dos nodos de la misma red se conectan entre sí en el caso de que ambas leyes fuesen aprobadas en un intervalo de tiempo prefijado. Se pueden agregar otras condiciones; como por ejemplo que las dos normas sean del mismo o diferente rango o que regulen el mismo u otro aspecto. En esta tesis se dan los primeros pasos en este sentido, aunque todavía hay un camino largo a recorrer.

Por analogía con los estudios QSAR/QSPR de moléculas, se puede intentar desarrollar modelos basados en LDA que discriminen las leyes más estables de otras que no lo son; o las concausas de un delito de la causa principal. En estos modelos S, sería una variable real de la red utilizada para cuantificar la estabilidad en el tiempo de la ley tributaria o la posibilidad de que una concausa aparente sea la causa principal. Además, por una parte, los $C_t(j)$ representarían centralidades de nodo (concausas criminales o leyes tributarias) de tipo t, seleccionado entre c clases de centralidades. Por otra parte, $a_0$ sería el término independiente y los $a_t$ serían los coeficientes de estas centralidades en el modelo QSPR jurídico con la forma linear:

$$S = \sum_{t=1}^{c} a_t \cdot C_t(j) + a_0 \quad (10)$$

Los parámetros estadísticos para la ecuación LDA serían esencialmente los mismos que en un QSAR: n = número de casos, U = Estadígrafo U o de λ de Wilks (U = 1 para discriminación perfecta y U = 0 para discriminación nula), F = Razón de Fisher, y p = el nivel de error. Podemos utilizar diferentes software para calcular TIs y/o $C_t(j)$ de estas redes, como los presentados en la **Tabla 3.**

Por ejemplo, a continuación presentamos por primera vez, un modelo QSPR no publicado aun que describe la dinámica del sistema tributario Español. La mejor ecuación QSPR encontrada fue:

$$S(L_{ij}) = 9.29 \cdot \left[ WC_1\left({}^{c}L_{ti}\right) - WC_1\left({}^{c}L_{ti+1}\right) \right] + 0.12 \quad (11)$$

$$n = 33{,}951 \quad \chi^2 = 32{,}920.85 \quad p < 0.001$$

Donde, $WC_k(L_{ti})$ y $WC_k(L_{ti+1})$ son centralidades de nodo de Rücker-Markov definidas por González-Díaz et al.; recientemente. Estos parámetros cuantifican la longitud de todos los caminos a partir de una ley dada $L_i$ con respecto a todas las $k^{th}$ leyes de tipo L, aprobadas antes y después en el sistema tributario español entre el tiempo $t_i$ y $t_{i+1}$. El modelo es capaz de reconstruir todo el record histórico de leyes tributarias con alta Especificidad (Es) y Sensibilidad (Sn), **Tabla 3**.

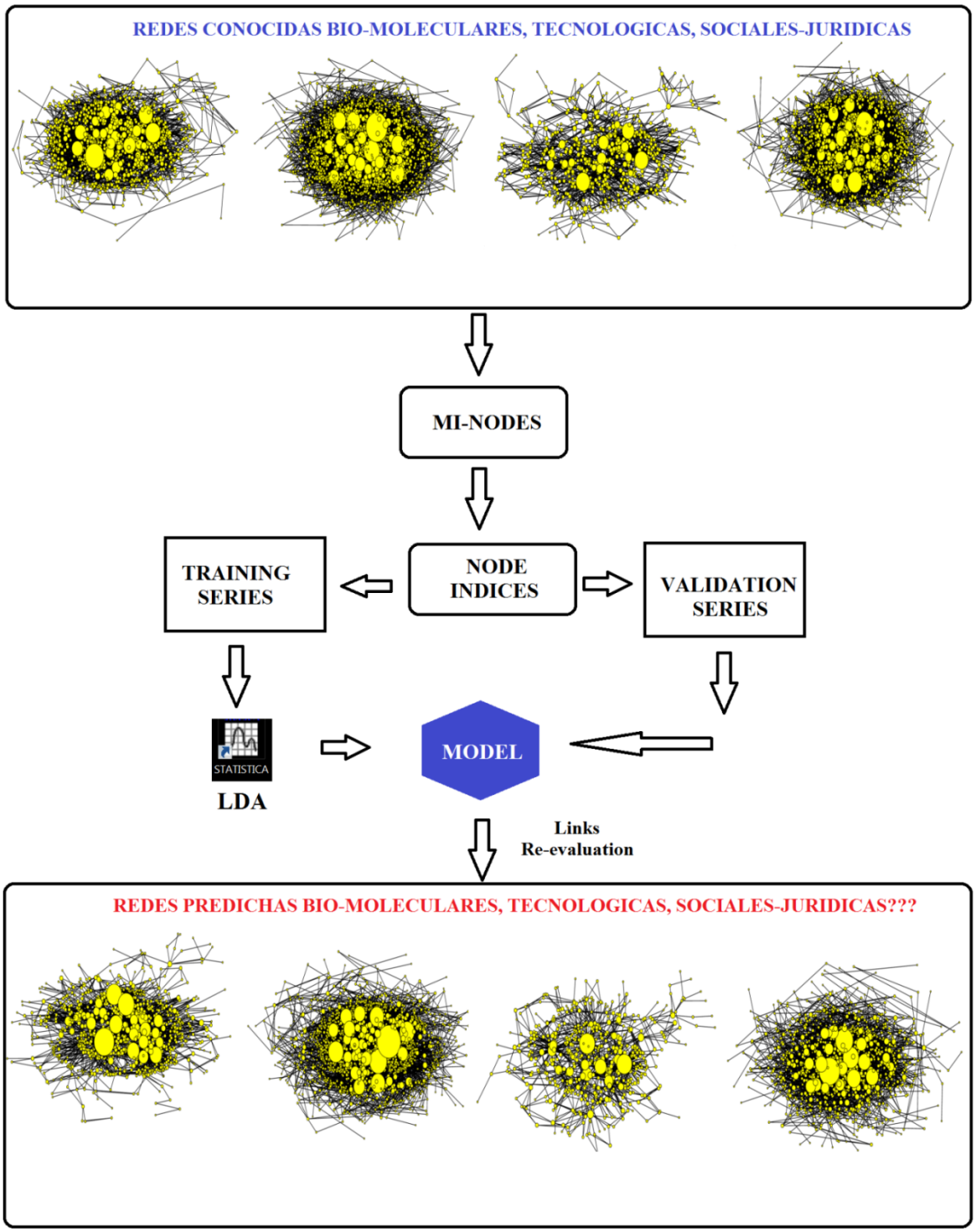



***Figura 9.*** *Diagrama de flujo general para los plazos principales dados en un estudio QSPR aplicado en SNA incluyendo redes jurídicas y otras.*

***Tabla 3.*** *Resultados del modelo predictivo del sistema tributario español.*

| Set de datos | Model Parameters | Classification results | | | |
|---|---|---|---|---|---|
| | | % | | NL | L |
| Entrenamiento | Es | 82.3 | NL | 17,597 | 3,910 |
| | Sn | 100 | L | 915 | 2,579 |
| | % | 90.1 | Total | | |
| Validación | Es | 81.8 | NL | 5,186 | 1,155 |
| | Sn | 100 | L | 0 | 5,014 |
| | % | 89.8 | Total | | |

*Filas: Leyes reales; Columnas: Leyes predichas. L = leyes que fueron emitidas en el intervalo de tiempo prefijado, NL = Leyes no emitidas en el intervalo de tiempo prefijado*

En la **Figura 10,** se muestra el grafo de la red jurídica que describe el sistema tributario español durante los años 1946-2004. En este caso, una de las aplicaciones más prometedoras también podría ser la búsqueda de modelos QSPR. Específicamente, modelos capaces de predecir, por ejemplo, que leyes serán más estables (tardan más en ser sustituidas) dada la representación grafo-teórica de todas las leyes anteriores. Todos los detalles, incluyendo valores de centralidad, tipo de ley, fecha en que fue aprobada, así como los ficheros de la red compleja, están disponibles si se solicitan al autor responsable de la correspondencia de este trabajo; aliuska.duardo@usc.es.

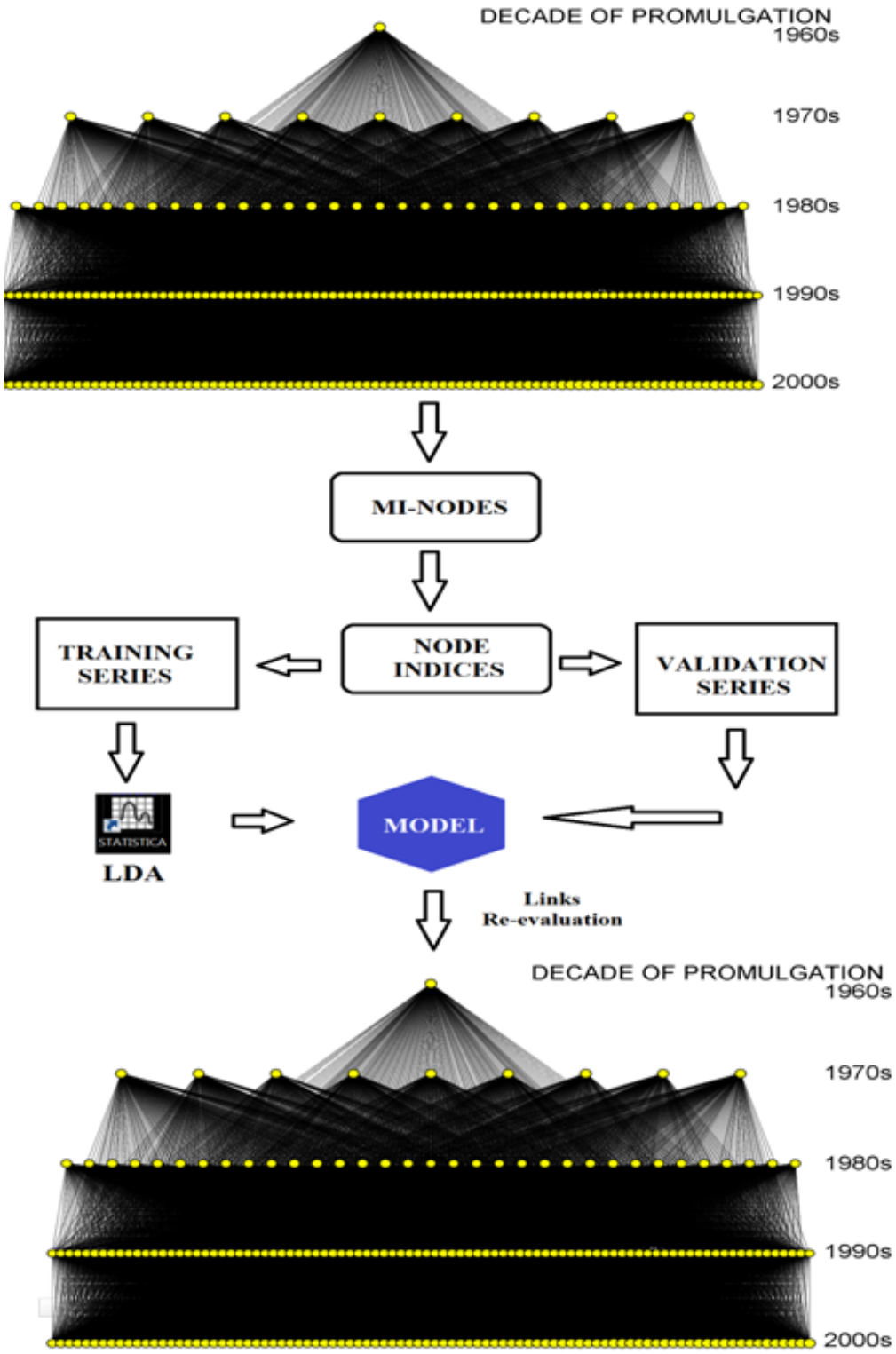



***Figura 10.*** *Representación en forma de red compleja y diagrama de un estudio QSPR de las leyes del sistema tributario Español.*

## Bibliografía


Devah, P. Mark of a Criminal Record. Am J Soc **2003**: 937-75.

Duardo-Sanchez, A. Criminal law networks, markov chains, Shannon entropy and artificial neural networks. In Complex Network Entropy: From Molecules to Biology, Parasitology, Technology, Social, Legal, and Neurosciences, González-Díaz, H.; Prado-Prado, F.J.; García-Mera, X., Eds. Transworld Research Network: Kerala, India, **2011**; pp 107-14.

Duardo-Sanchez, A. Study of criminal law networks with Markov-probability centralities. In Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks, González-Díaz, H.; Munteanu, C.R., Eds. Transworld Research Network: Kerala, India, **2010**; pp 205-12.

González-Díaz H, Riera-Fernández P, Pazos A, Munteanu CR. The Rücker-Markov invariants of complex Bio-Systems: Applications in Parasitology and Neuroinformatics. Biosystems. **2013**; 111(3):199-207.

Gonzalez-Diaz, H.; Riera-Fernandez, P., New Markov-autocorrelation indices for re-evaluation of links in chemical and biological complex networks used in metabolomics, parasitology, neurosciences, and epidemiology. J Chem Inf Model **2012,** 52 (12), 3331-40.

Riera-Fernández P, Munteanu CR, Escobar M, Prado-Prado F, Martín-Romalde R, Pereira D, Villalba K, Duardo-Sánchez A, González-Díaz H. New Markov-Shannon Entropy models to assess connectivity quality in complex networks: from molecular to cellular pathway, Parasite-Host, Neural, Industry, and Legal-Social networks. J Theor Biol. **2012**; 293:174-88.

Riera-Fernandez, P.; Munteanu, C. R.; Dorado, J.; Martin-Romalde, R.; Duardo-Sanchez, A.; Gonzalez-Diaz, H., From Chemical Graphs in Computer-Aided Drug Design to General Markov-Galvez Indices of Drug-Target, Proteome, Drug-Parasitic Disease, Technological, and Social-Legal Networks. Current Computer-Aided Drug Design **2011,** 7 (4), 315-337.

Riera-Fernandez, P.; Munteanu, C. R.; Escobar, M.; Prado-Prado, F.; Martin-Romalde, R.; Pereira, D.; Villalba, K.; Duardo-Sanchez, A.; Gonzalez-Diaz, H., New Markov-Shannon Entropy models to assess connectivity quality in complex networks: from molecular to cellular pathway, Parasite-Host, Neural, Industry, and Legal-Social networks. J Theor Biol **2012,** 293, 174-88.

Riera-Fernandez, P.; Munteanu, C. R.; Pedreira-Souto, N.; Martin-Romalde, R.; Duardo-Sanchez, A.; Gonzalez-Diaz, 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.

## II.7. Aspectos jurídicos relacionados con las TICs

*"Society must worry about both long-term implications of the hypothetical future and short-term realities of the present." Ingfei Chen*

*"Biotechnology is nothing new, except for lawyers". J. Davies*

Existen múltiples interacciones entre la Biotecnología, la Bioinformática, las Ciencias Biomédicas y el Derecho. Esta interacción, en el caso de las TICs y el Derecho, comprende tanto la protección de los programas científicos desarrollados en este marco como la validación y protección de los propios resultados obtenidos tras su utilización. En efecto, la relación entre la Bioinformática y el mundo de lo jurídico abarca una amplia gama de temas diferentes dónde se incluye: la propiedad intelectual (marcas, secretos industriales, etc.), las licencias (de patentes, de explotación, de propiedad intelectual de software, etc.), la legislación y la reglamentación del desarrollo de productos; así como cuestiones jurídico-corporativas, incluso concernientes a cuestiones ético-legales relacionadas con la patentabilidad de las formas vivas, o la pertinencia y la disponibilidad de las plantas transgénicas. Estos son algunos de los temas que se discuten en la parte del trabajo experimental de la presente tesis. Especialmente se hace referencia a la validación de los modelos QSAR con aplicaciones en el descubrimiento de medicamentos; la protección jurídica del software científico; la patentabilidad de los genomas de plantas, o la tributación de los programas de ordenador.

### Bibliografía

Barratt, R.; Balls, M. An overall strategy for the testing of chemicals for human hazard and risk assessment under the EU REACH system. ATLA **2003**; 31: 19-20.

Bastian, M.; Hetmann, S.; Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. In International AAAI Conference on Weblogs and Social Media (ICWSM09), North America, 2009.

Batagelj, V.; Mrvar, A. Pajek 1.15. **2006**.

Benfenati, E. The CAESAR project for in silico models for the REACH legislation. Chem Cent J **2010**; 4 Suppl 1: I1.

Benfenati, E. The specificity of the QSAR models for regulatory purposes: the example of the DEMETRA project. SAR & QSAR in Environmental Research **2007**; 18: 209-20.

Berca, M.N.; Duardo-Sanchez, A.; González-Díaz, H.; Pazos, A.; Munteanu, C.R. Markov entropy for biology, parasitology, linguistic, technology, social and law networks. In Complex Network Entropy: From Molecules to Biology, Parasitology, Technology, Social, Legal, and Neurosciences, González-Díaz, H.; Prado-Prado, F.J.; García-Mera, X., Eds. Transworld Research Network: Kerala, India, 2011; pp 127-42.

Bouyssie, D.; Gonzalez de Peredo, A.; Mouton, E.; Albigot, R.; Roussel, L.; Ortega, N.; Cayrol, C.; Burlet-Schiltz, O.; Girard, J.P.; Monsarrat, B. Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: application to the proteomics study of membrane proteins from primary human endothelial cells. Mol Cell Proteomics **2007**; 6: 1621-37.

Coecke, S.; Ahr, H.; Blaauboer, B.J. Metabolism: a bottleneck in in vitro toxicological test development. The Report and Recommendations of ECVAM Workshop 54. ATLA **2006**; 34: 49-8.

Cronin, M.T.; Jaworska, J.S.; Walker, J.D.; Comber, M.H.; Watts, C.D.; Worth, A.P. Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. Environ Health Perspect **2003**; 111: 1391-401.

Davidson, S.J.; Bergs, S.J.; Kapsner, M. Open, click, download, send … What have you agreed to? The possibilities seem endless. Computer Associations Law Conference (CLA 2001). **2001**.

Duardo-Sánchez, A; Patlewicz, G; López-Díaz, A. Current Topics on Software Use in Medicinal Chemistry: Intellectual Property, Taxes, and Regulatory Issues. Current Topics in Medicinal Chemistry, **2008**, 8(18).

Edler, L.; Poirier, K.; Dourson, M.; Kleiner, J.; Mileson, B.; Nordmann, H.; Renwick, A.B.; Slob, W.; Walton, K.; Wurtzen, G. Mathematical modeling and quantitative methods. Food and Chemical Toxicology **2002**; 40: 283–326.

Fjodorova, N.; Novich, M.; Vrachko, M.; Kharchevnikova, N.; Zholdakova, Z.; Sinitsyna, O.; Benfenati, E. Regulatory assessment of chemicals within OECD member countries, EU and in Russia. J Environ Sci Health C Environ Carcinog Ecotoxicol Rev **2008**; 26: 40-88.

Fjodorova, N.; Novich, M.; Vrachko, M.; Smirnov, V.; Kharchevnikova, N.; Zholdakova, Z.; Novikov, S.; Skvortsova, N.; Filimonov, D.; Poroikov, V.; Benfenati, E. Directions in QSAR modeling for regulatory uses in OECD member countries, EU and in Russia. J Environ Sci Health C Environ Carcinog Ecotoxicol Rev **2008**; 26: 201-36.

Hem, E.; Bordahl, P.E. Max Sanger - father of the modern caesarean section. Gynecologic & Obstetric Investigation **2003**; 55: 127-9.

International Legal Protection for Software. http://www.softwareprotection.com/ (25/9/2007).

Jansen, F.K.; Freytag, G. Immune reactions to fractions of crystalline insulin. II. May peri-insulitis be produced by an antigen different from true Sanger insulin? Diabetologia **1973**; 9: 191-6.

Jaworska, J.; Comber, M.; Auer, C.; Van Leeuwen, C.J. Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. Environmental Health Perspectives Mini Monografo **2003**; 10: 1358–60.

Koeter, H.B.; Visser, R. Work in OECD on chemical safety: approaches for human risk assessment. Industrial Health **2000**; 38: 109-19.

Morcon, C.; Roughton, A.; Gaham, J. The Modern Law of Trade Marks. Butterworth: London 2005.

OECD. Report on the Regulatory Uses and Applications in OECD Member Countries of (Quantitative) Structure-Activity Relationchip ((Q)SAR) Models in the Assessment of New and Existing Chemicals. In OECD Environment Health and Safety Publications: Paris, 2006.

Porcelli, C.; Boriani, E.; Roncaglioni, A.; Chana, A.; Benfenati, E. Regulatory perspectives in the use and validation of QSAR. A case study: DEMETRA model for Daphnia toxicity. Environmental Science and Technology **2008**; 42: 491-6.

Porcelli, C.; Roncaglioni, A.; Chana, A.; Benfenati, E. A comparison of DEMETRA individual QSARs with an index for evaluation of uncertainty. Chemosphere **2008**; 71: 1845-52.

Richards, T.W.; Peirce, B.O.; Baxter, G.P. Charles Robert Sanger. Science **1912**; 35: 532.

Rowland, D.; Campbell, A. Supply of Software: Copyright and Contract Issues. International Journal of Law and Information Technolog **2002**; 10: 23-40(18).

Rudnick, P.A.; Wang, Y.; Evans, E.; Lee, C.S.; Balgley, B.M. Large scale analysis of MASCOT results using a Mass Accuracy-based THreshold (MATH) effectively improves data interpretation. J Proteome Res **2005**; 4: 1353-60.

Steering Committee for Intellectual Property Issues in Software Computer Science and Telecommunications Board Commission on Physical Sciences, M., and Applications National Research Council. Intellectual Property Issues In software. National Academy Press: Washington, D.C 1991.

Story, A. Intellectual Property and Computer Software. In Intellectual Property Rights and Sustainable Development, ICTSD-UNCTAD, Ed. Imprimerie Typhon: Chavanod, 2004; p 12.

Walker, J.; Jaworska, J.; Comber, M.; Schultz, T.; Dearden, J. Guidelines for developing and using quantitative structure–activity relationships. Environmental Toxicology and Chemistry **2003**; 22: 1653–65.

Westkamp, G.N. Protección del material biológico mediante derechos de autor. ¿Vuelta de la Bioinformática a los Derechos de autor en la biotecnología? IPR-Helpdesk Bulletin **2005**.

WIPO. Madrid Agreement Concerning the International Registration of Marks of April 14, 1891, as revised at Brussels on December 14, 1900, at Washington on June 2, 1911, at The Hague on November 6, 1925, at London on June 2, 1934, at Nice on June 15, 1957, and at Stockholm on July 14, 1967,1 and as amended on September 28, **1979**.

WIPO. Protocol Relating to the Madrid Agreement Concerning the International Registration of Marks adopted at Madrid on June 27, 1989 and amended on October 3, **2006**.

Worth, A.P.; Hartung, T.; Van Leeuwen, C.J. The role of the European centre for the validation of alternative methods (ECVAM) in the validation of (Q)SARs. SAR & QSAR in Environmental Research **2004**; 15: 345-58.

Worth, A.P.; Van Leeuwen, C.J.; Hartung, T. The prospects for using (Q)SARs in a changing political environment-high expectations and a key role for the European Commission's joint research centre. SAR & QSAR in Environmental Research **2004**; 15: 331-43.

Yang, C.G.; Granite, S.J.; Van Eyk, J.E.; Winslow, R.L. MASCOT HTML and XML parser: an implementation of a novel object model for protein identification data. Proteomics **2006**; 6: 5688-93.

# III. TRABAJO EXPERIMENTAL

## Modelos QSPR de Problemas Bio-Moleculares y Jurídicos

## III.1. Nuevo software para índices topológicos de Markov de redes complejas

**Artículo 1:** S2SNet: A Tool for Transforming Characters and Numeric Sequences into Star Network Topological Indices in Chemoinformatics, Bioinformatics, Biomedical, and Social-Legal Sciences. C.R. Munteanu, A.L. Magalhaes, **A. Duardo-Sánchez**, A. Pazos, and H. González-Díaz. *Current Bioinformatics*, **2013**, 8, 429-437.

El estudio de una gran variedad de sistemas desde las redes de proteínas o sistemas complejos representados por redes jurídico-sociales puede llevarse a cabo con la teoría de redes complejas. Es de particular interés la cuantificación numérica de información significativa contenida por las secuencias de aminoácidos (proteínas), nucleótidos (genes), señales eléctricas (electroencefalogramas o electrocardiogramas) o en textos legales o series de frecuencia de uso de distintos tipos de leyes fiscales. En este trabajo, se describe y se lleva a cabo una revisión del software S2SNet. Una herramienta desarrollada por CR Munteanu y H González-Díaz en trabajos previos. La nueva herramienta implementada en el lenguaje de programación *"Python",* con un interfaz de usuario como el que se puede ver en la **Figura 11-a**, puede transformar/representar cualquier secuencia de caracteres, o números en serie, en grafos con topología de tipo estrella. Los nodos del grafo son las partes del sistema (aminoácidos, nucleótidos, organismos, empresas, poblaciones, leyes). Las ramas del grafo con topología estrella son sub-grafos lineales (todos los nodos tienen grado $\leq 2$). El software permite construir grafos de recurrencia donde la secuencia de nodos se incorpora *"embeded graphs*". A partir de estos grafos, se pueden clacular TIs para hacer estudios QSPR. En la **Figura 11-b**, se muestran ambos tipos de grafos para el sistema de tributario de leyes de España.

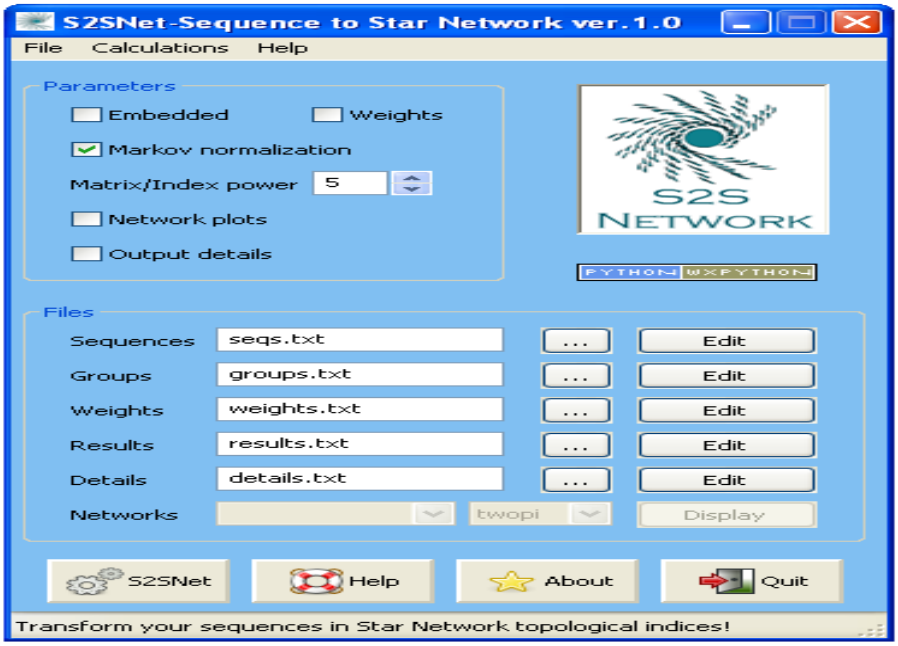



***Figura 11-a.*** *S2SNet: interface usuario - software*

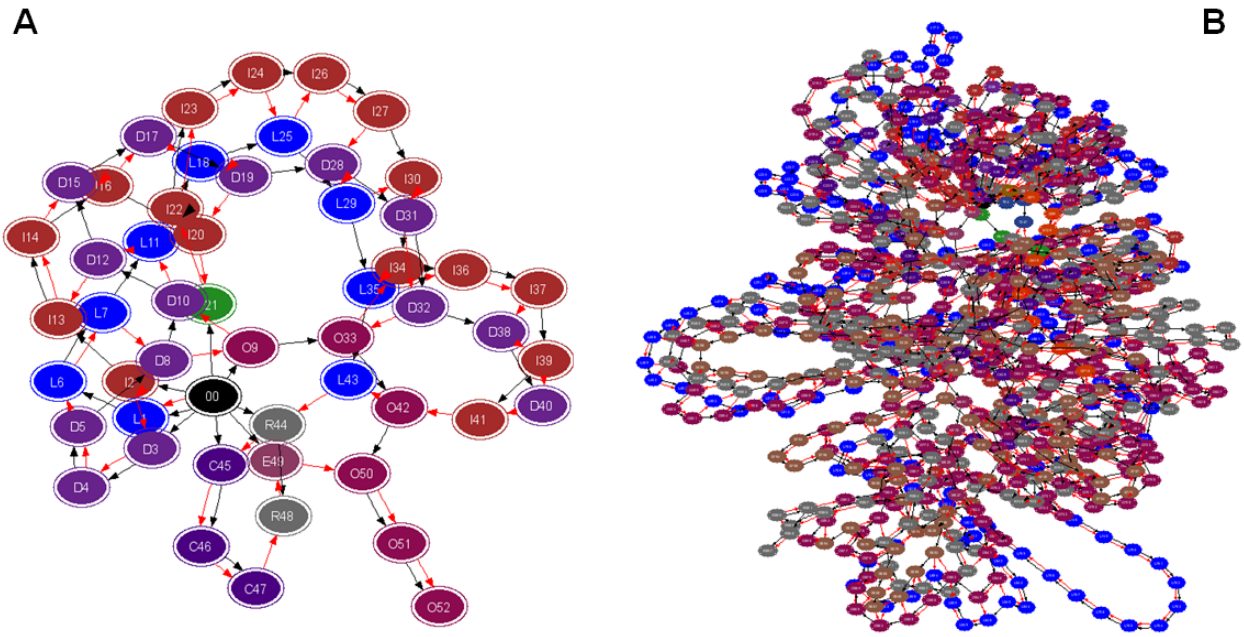



***Figura 11-b***. *Red para el sistema tributario español en dos periodos: A) 1946 - 1978 y B) 1946 - 2004 (comportamiento acumulativo con respect a 1946).*

## III.2. Predicción de Redes Complejas con Índices Markov-Wiener y ANNs

**Artículo 2:** Modeling Complex Metabolic Reactions, Ecological Systems, and Financial and Legal Networks with MIANN Models Based on Markov-Wiener Node Descriptors. **A. Duardo-Sánchez**, C.R. Munteanu, P. Riera-Fernández, A. López-Díaz, A. Pazos, and H. González-Díaz. *Journal of Chemical Information and Modelling*, **2014**, 54, 16-29

En 1947, Wiener publicó un artículo en el cual propuso que las propiedades físicas de los compuestos orgánicos dependen funcinalmente del número, clase, la organización estructural de los átomos en las moléculas. La ecuación usada por Wiener puede ser representada con la siguiente fórmula [5-7]:

$$\Delta t = \frac{a}{n^2} \cdot \Delta W + b \cdot \Delta p \qquad (1) \qquad\qquad W = \frac{1}{2} \cdot \sum_{i=1}^{D} \sum_{j=1}^{D} d_{ij} \qquad (2)$$

Donde, n es el número de átomos, *p* es el número de polaridad definido como el número de pares de átomos de carbono que están separados por tres enlaces y *W* es el número de caminos; definido como las suma de las distancias topológicas entre dos átomos $d_{ij}$ en la matriz de distancias (**D**). Este parámetro es considerado uno de los TIs más antiguos. En 1971 Hosoya acuñó el término para referirse al índice Z y es usado actualmente para definir índice numérico de la topología molecular. Los TIs se derivan matemáticamente de los grafos estructurales de las moléculas usualmente construidos con Hydrógenos suprimidos (H-depleted molecular graph). El índice de caminos es también llamado Índice de Wiener ó número de Wiener (W). Como puede verse, los pares de nodos más distantes tienen una contribución más alta al índice que los más adyacentes. Es interesante apuntar que el índice de Wiener fue propuesto de manera independiente en 1959 por Harary, en el contexto de la sociometría, con el nombre de estado total del grafo (*total status of a graph*). Así mismo, también fue discutido por Rouvray and Crafford en 1975. Esto indica que W no era bien conocido en esta época. Sin embargo, a mediados de los 70s muchos autores comienzan a estudiar las

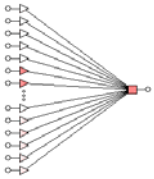
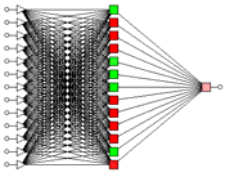
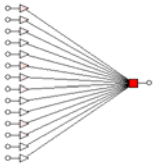
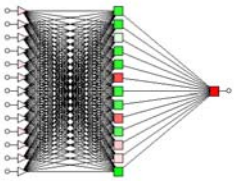
propiedades y aplicaciones de los TIs. Lo cuál llevó al desarrollo de nuevos TIs, algunos de ellos basados en W.

En este trabajo, se introducen las centralidades de Markov-Wiener de orden k-ésimo ($W_k$) usando las probabilidades de interacción entre vértices ($p_{ij}$) basadas en la k-ésima potencia ($^1\mathbf{\Pi}$)$^k$ de la matriz estocástica $^1\mathbf{\Pi}$. Se podrán sumar todos los nodos del grafo o solo los nodo (j) conectados al átomo (i); condición representada por $j \rightarrow i$, cuya suma es el grado de este nodo ($\delta_i$):

$$W_k(G) = \frac{1}{2} \cdot \sum_{i=1}^{D} \sum_{j=i}^{D} {}^k p_{ij} \cdot d_{ij} \qquad (3) \qquad\qquad W_k(i) = \frac{1}{2} \cdot \sum_{i=1}^{1} \sum_{j \rightarrow i}^{\delta_i} {}^k p_{ij} \cdot d_{ij} = \frac{1}{2} \cdot \sum_{j \rightarrow i}^{\delta_i} {}^k p_{ij} \cdot d_{ij} \qquad (4)$$

Es preciso no confundir estos índices con los procesos estocásticos de Markov-Wiener. En este trabajo se calculan los valores de $W_k(i)$ para >100,000 nodos en >100 redes complejas con el programa MI-NODES. Incluyendo: Redes de Reacciones Metabólicas (RRM) de 40 organismos, 75 redes ecológicas de la base de datos Interaction Web Database Biological Networks (IWDBNs) y el Sistema Tributario Español desde 1946-2004. Los valores de $W_k(i)$ fueron usados como entradas para entrenar y validar diferentes ANNs con el objetivo de discriminar patrones correctos de conectividad. Los modelos MIANN obtenidos tienen valores prometedores de Sensibilidad y Especificidad (Sn = Sensitivity / Specificity = Sp (ver **Tabla 4**).

***Tabla 4.*** *Modelos ANN con índices Markov-Wiener para redes complejas*

| Redes de Reacciones Metabólicas | $L_i$ | Li = 1 | Li = 0 | % | Pr. | % | Li = 1 | Li = 0 |
|---|---|---|---|---|---|---|---|---|
| LNN 15:15-1:1 | Li = 1 | 7276 | 1985 | 78.1 | Sn | 77.9 | 21917 | 6156 |
|  | Li = 0 | 2044 | 7066 | 78.1 | Sp | 77.6 | 6227 | 21329 |
| Redes de Ecosistemas | Li | Li = 1 | Li = 0 | % | Pr. | % | Li = 1 | Li = 0 |
| MLP 13:13-13-1:1 | Li = 1 | 4570 | 547 | 91.1 | Sn | 90.5 | 1363 | 194 |
|  | Li = 0 | 449 | 4346 | 88.8 | Sp | 88.1 | 143 | 1437 |
| Sistema Legal Financiero Español | Li | Li = 1 | Li = 0 | % | Pr. | % | Li = 1 | Li = 0 |
| LNN 14:14-1:1 | Li = 1 | 125 | 41 | 86.2 | Sn | 87.4 | 370 | 156 |
|  | Li = 0 | 18 | 298 | 85.4 | Sp | 87.9 | 59 | 914 |
| MLP 14:14-14:1 | Li = 1 | 119 | 54 | 85.3 | Sn | 83.2 | 366 | 129 |
|  | Li = 0 | 24 | 285 | 87.9 | Sp | 84.1 | 63 | 941 |

## III.3. Predicción de Redes Complejas con Índices Markov-Balaban y ANNs

**Artículo 3:** MIANN Models of Networks of Biochemical Reactions, Ecosystems, and U.S. Supreme Court with Balaban-Markov Indices. **A. Duardo**, H. González-Díaz, and A Pazos. *Current Bioinformatics*, **2014**, 9, *en imprenta*.

El Prof. Alexandru T. Balaban introdujo un de los TIs más famosos que en la actualidad es conocido como el índice J de Balaban (Balaban's J index) [14]. El índice J depende de q = número de aristas del grafo molecular, n = número de nodos (átomos) y μ = (q – n + 1) = número ciclomático. El índice J también depende de $S_i$ = suma de las distancias por la i-ésima fila ó j-ésima columna de la matriz de distancias topológicas **D** del grafo molecular con Hydrógenos suprimidos. Su formula es:

$$J(G) = \frac{q}{\mu + 1} \cdot \sum_{edges}^{q} \left( S_i \cdot S_j \right)^{-1/2} \qquad (1)$$

En este trabajo, se usan las cadenas de Markov para generalizar el índice J y aplicarlo a la bioinformática, la biología de sistemas y las ciencias sociales. También se ofrece la definición de índices Markov-Balaban $J_k(G)$ totales y las centralidades $J_k(i)$ para los nodos de las redes complejas basadas en esta idea:

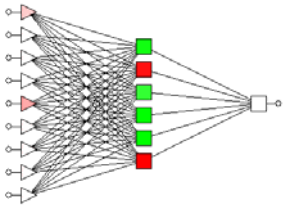
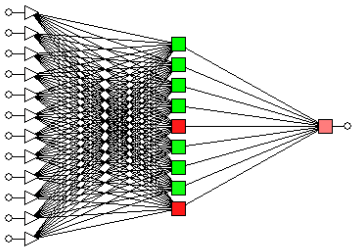
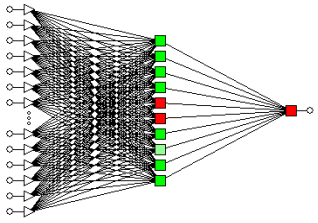
$$J_k(G) = \frac{q}{\mu + 1} \cdot \sum_{edges}^{q} \left( {}^{k}p_{ij} \cdot S_i \cdot S_j \right)^{-1/2} \quad (2)$$

$$J_k(i) = \frac{q}{\mu + 1} \cdot \sum_{i-edges}^{\delta_i} \left( {}^{k}p_{ij} \cdot S_i \cdot S_j \right)^{-1/2} \quad (3)$$

Se buscan nuevos modelos ANN para mostrar la capacidad de discriminación de los nuevos índices $J_k(i)$ a nivel de nodo en tres experimentos. En primer lugar, se calcularon más de 1.000.000 de valores de $J_k(i)$ y otros índices para todos los nodos de >100 redes complejas. En los tres experimentos, encontramos nuevos modelos MIANN con > 80 % de especificidad (Sp) y sensibilidad (Sn) en series de entrenamiento y validación externa (ver **Tabla 5**). Se usaron Redes de Reacciones Metabólicas de 42 organismos (bacterias, hongos, nematodos y plantas), 73 Redes

Ecológicas y 43 sub-redes de citas de casos de la Corte Suprema de Estados Unidos en diferentes décadas entre 1791 y 2005. Este trabajo puede abrir una nueva ruta para la aplicación de los TI tipo Balaban para desentrañar relaciones ocultas entre estructura y propiedades de las redes complejas bio-moleculares, ecológicas y jurídico-sociales.

***Tabla 5.*** *Modelos ANN con indices Markov-Balaban para distintas redes complejas*

| Redes Reacciones Metabólicas | Li | Li = 1 | Li = 0 | % | Pr. | % | Li = 1 | Li = 0 |
|---|---|---|---|---|---|---|---|---|
| MLP 9:9-6-1:1  | Li = 1 | 29117 | 19490 | 81.8 | Sn | 81.9 | 9729 | 6426 |
| | Li = 0 | 6481 | 87304 | 81.7 | Sp | 81.9 | 2137 | 29172 |
| Redes de Ecosistemas | Li | Li = 1 | Li = 0 | % | Pr. | % | Li = 1 | Li = 0 |
| MLP 12:12-9-1:1  | Li = 1 | 3992 | 2684 | 81.6 | Sn | 81.3 | 1326 | 898 |
| | Li = 0 | 902 | 11934 | 81.6 | Sp | 81.6 | 305 | 3975 |
| Corte Suprema de EE.UU. | Li | Li = 1 | Li = 0 | % | Pr. | % | Li = 1 | Li = 0 |
| MLP 18:18-10-1:1  | Li = 1 | 81225 | 51008 | 82.49 | Sn | 82.76 | 26985 | 17014 |
| | Li = 0 | 16917 | 243415 | 82.66 | Sp | 82.7 | 5728 | 81128 |

## III.4. Predicción de Redes Bio-Moleculares, Socio-Económicas y Jurídicas

**Artículo 4:** MI-NODES Multiscale Models of Metabolic Reactions, Ecological, Brain Connectome, Epidemic, World Trade, and Legal-Social Networks. **A. Duardo**, H. González-Díaz, and A Pazos. *Current Bioinformatics*, **2014**, 9, *en imprenta*.

Como se ha visto a lo largo de esta tesis, los sistemas y redes complejas aparecen en casi todos los ámbitos de la realidad. Se encuentran desde las redes de residuos (aminoácidos) en la estructura 3D de las proteínas, hasta una escala superior en las redes de interacción entre proteínas (PINs). Las reacciones químicas forman Redes de Reacciones Metabólicas (RRMs) en organismos vivos, pero también forman redes de reacciones atmosféricas en planetas y lunas. Aparecen Redes de Neuronas en el gusano *C. elegans*, en el conectoma del Cerebro humano, o en las Redes de Neuronas Artificiales (ANNs).

Existen redes epidemiológicas de propagación de infecciónes por brotes contagiosos en los seres humanos y las redes informáticas (Internet o las redes inalámbricas) debido a la infección por malware con un software viral. Redes sociales regidas por diferentes reglas evolucionaron a partir de la inteligencia de enjambre hasta las redes de las sociedades de cazadores - recolectores, o redes socio-jurídicas reflejo de sociedades complejas como la red de la Corte Suprema de los EE.UU.

En todos estos casos, se puede hacer la misma pregunta. ¿Se pueden predecir vínculos o conexiones en redes complejas a partir de la información estructural? En este trabajo de revisión y la tesis se ha abordado este problema usando las mismas ideas empleadas de uso común en quimio-informática en estudios cuantitativos de estructura – propiedad; más conocidos en idioma Íngles como técnicas de tipo Quantitative Structure-Property Relationships (QSPR).

Para esto se ha usado la teoría de cadenas de Markov para generalizar TIs clásicos, se han implementado en los software MI-NODES / MARCH-INSIDE y se han usado como inputs de ANNs para obtener modelos definidos por nosotros como MIANN.

También en este trabajo se ha realizado una revisión de modelos de este tipo basados en ANNs u otros clasificadores y con otros índices desarrollados anteriormente por nuestro grupo.

En la **Tabla 6.**, se ilustran los resultados econtrados con modelos MIANN y otros lineares para distintos tipos de redes complejas y TIs de tipo Markov calculados con MI-NODES. En el centro de esta tabla también se ilustra el diagrama de flujo para los pasos generales que se deben dar en estudios QSPR de redes complejas.

***Tabla 6.*** *Distintas aplicaciones de modelos MI-NODES*

| Red[a] | Par. | $^{k}C_{BM}(j)$ | $^{k}C_{\theta}(j)^{k}_{i}$ | $^{k}C_{\pi}(j)^{k}_{i}$ | $^{k}C_{wc}(j)$ | $^{k}C_{\chi}(j)$ |
|---|---|---|---|---|---|---|
| Train | | | | Train | | |
| RM | Sp | 72.22 | 99.98 | ? | 81.32 | 70.19 |
| | Sn | 71.25 | 87.24 | ? | 73.91 | 70.63 |
| PH | Sp | 87.49 | 95.4 | 87.49 | 95.24 | 90.56 |
| | Sn | 100 | 72.22 | 100 | 73.27 | 92.70 |
| CC | Sp | 84.14 | 92.2 | 98.49 | 88.40 | 75.32 |
| | Sn | 72.70 | 71.2 | 73.30 | 74.64 | 94.69 |
| SF | Sp | 87.14 | 99.2 | 93.21 | 71.49 | 100 |
| | Sn | 72.68 | 70.4 | 72.01 | 71.64 | 89.70 |
| Validation | | | | Validation | | |
| RM | Sp | 72.28 | 99.96 | ? | 81.82 | 71.17 |
| | Sn | 71.24 | 86.91 | ? | 73.81 | 70.89 |
| PH | Sp | 87.67 | 95.5 | 87.67 | 95.43 | 91.00 |
| | Sn | 100 | 72 | 100 | 70.81 | 92.83 |
| CC | Sp | 84.42 | 92.5 | 98.41 | 88.30 | 75.51 |
| | Sn | 71.88 | 70.4 | 71.21 | 73.27 | 94.73 |
| SF | Sp | 87.34 | 99.1 | 93.20 | 71.55 | 100 |
| | Sn | 75.78 | 74.2 | 73.47 | 70.54 | 90.22 |

[a] *Red de Reacciones Metabólicas (RM), Parsito-Huésped (PH), Corteza Cerebral (CC) y Sistema Financiero (SF) de leyes.*

# IV. CONCLUSIONES

## IV. Conclusiones

4.1. Los modelos QSPR basados en ANNs y TIs de grafos, comúnmente empleados en Ciencias Bio-Moleculares, pueden extenderse a las Ciencias Jurídico-Sociales.

4.2. Los TIs de utilidad en Ciencias Jurídico-Sociales pueden calcularse a partir de redes de nodos que pueden representar aspectos tan diferentes como leyes tributarias o concausas penales.

4.3. La teoría de las Cadenas de Markov puede usarse para generalizar TIs clásicos creando TIs estoscásticos análogos de orden k.

4.4. Los TIs basados en Cadenas de Markov pueden ser usados como variables de entradas (inputs) de ANNs para predecir redes complejas.

4.5. El cálculo de los nuevos TIs estocásticos pueden ser implementados en programas computacionales con una interface usuario-ordenador que facilite su uso por investigadores no expertos en informática.

4.6. Los nuevos programas S2SNet y MI-NODES, revisados en esta tesis, pueden convertirse en importantes herramientas para el cálculo de TIs de distintos tipos de redes.

4.7. Podemos combinar operdaros de tipo Box-Jenkins usados en series de tiempo con TIs para encontrar modelos multi-objetivo QSPR-ANN de redes complejas.

4.8. Los TIs de tipo Markov-Balaban y Markov-Wiener introducidos desmostraron ser útiles en estudios QSPR-ANN de redes complejas incluyendo redes Bio-Moleculares (Redes de Reacciones Metabólicas de diferentes organismos y Redes de diferentes Ecosistemas) así como redes Jurídico-Sociales (Red del Sistema Legal Financiero Español, y Red de Decisiones de la Corte Suprema de USA).

4.9. A pesar de existir todo un marco jurídico-legal para la protección de resultados científicos obtenidos a partir de la aplicación de las TICs en Ciencias Bio-Moleculares se puede avanzar aun más en la regulación en este sentido.

# V. FUTUROS DESARROLLOS

## V. FUTUROS DESARROLLOS

4.1. Extender los modelos QSPR basados en ANNs y TIs de grafos a redes complejas que integren información de Ciencias Bio-Moleculares y Ciencias Jurídico-Sociales conjuntamente.

4.2. Aplicar los modelos QSPR basados en ANNs y TIs de utilidad en Ciencias Jurídico-Sociales para representar / analizar otras redes como decisiones del Senado y Congreso de USA, Evolución de la Presión Fiscal en países de la OECD, *etc*.

4.3. Combinar los TIs basados en Cadenas de Markov con otros operadores diferentes de los de Box-Jenkins para encontrar modelos multi-objetivo QSPR-ANN de redes complejas.

4.4. Definir otras familias de TIs basados en Cadenas de Markov y TIs clásicos y explorar au utilidad en estudios QSPR-ANN de redes complejas incluyendo redes Bio-Moleculares y redes Jurídico-Sociales.

# VI. ANEXOS

## VI.1. ÍNDICE DE ABREVIATURAS

ANN – Artificial Neural Network = Red Neuronal Artificial

LDA – Linear Discriminant Analysis = ADL – Análisis Discriminante Linear

QSAR – Quantitative Structure-Activity Relationship = Relación Cuantitativa-Estructura Actividad

QSPR – Quantitative Structure-Property Relationship = Relación Cuantitativa-Estructura Propiedad

QSTR – Quantitative Structure-Toxicity Relationship = Relación Cuantitativa-Estructura Toxicidad

SNA – Social Networks Analysis = ARS – Análisis de Redes Sociales

TICs – Tecnologías de la Información y las Comunicaciones

TIs – Topological Indices = Índices Topológicos

## VI.2. ÍNDICE DE TABLAS

## VI.3. ÍNDICE DE FIGURAS

## VI.4. PUBLICACIONES: ARTÍCULOS PUBLICADOS POR LA DOCTORANDA EN EL TEMA DE LA TESIS

IV.1. *Current Bioinformutics*, **2013**, 8, 429-437.

IV.2. *Journal of Chemical Information and Modelling*, **2014**, 54, 16-29.

IV.3. *Current Bioinformatics*, **2014**, 9, *en imprenta*.

V.4. *Current Bioinformatics*, **2014**, 9, *en imprenta.*

**Artículo 1.**

**Nuevo software para índices topológicos de Markov de redes complejas**

S2SNet: A Tool for Transforming Characters and Numeric Sequences into Star Network Topological Indices in Chemoinformatics, Bioinformatics, Biomedical, and Social-Legal Sciences. C.R. Munteanu, A.L. Magalhaes, A. Duardo-Sánchez, A. Pazos, and H. González-Díaz. *Current Bioinformutics*, **2013**, 8, 429-437.

*Current Bioinformatics*, 2013, *8*, 429-437 

# S2SNet: A Tool for Transforming Characters and Numeric Sequences into Star Network Topological Indices in Chemoinformatics, Bioinformatics, Biomedical, and Social-Legal Sciences

Cristian R. Munteanu[*,1], Alexandre L. Magalhães[2], Aliuska Duardo-Sánchez[1,3], Alejandro Pazos[1] and Humberto González-Díaz[4,5]

[1]*Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruña, 15071 A Coruña, Spain*

[2]*REQUIMTE/University of Porto, Faculty of Science, Chemistry Department, 4169-007 Porto, Portugal*

[3]*Department of Special Public Law, Financial and Tributary Law Area, Faculty of Law, University of Santiago de Compostela (USC), 15782, Spain*

[4]*Department of Organic Chemistry II, University of the Basque Country, UPV/EHU, 48940, Leioa, Spain*

[5]*IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain*

**Abstract:** The study of complex systems such as proteins/DNA/RNA or dynamics of tax law systems can be carried out with the complex network theory. This allows the numerical quantification of the significant information contained by the sequences of amino acids, nucleotides or types of tax laws. In this paper we describe S2SNet, a new Python tool with a graphical user interface that can transform any sequence of characters or numbers into series of invariant star network topological indices. The application is based on Python reusable processing procedures that perform different functions such as reading sequence data, transforming numerical series into character sequences, changing letter codification of strings and drawing the star networks of each sequence using Graphviz package as graphical back-end. S2SNet was previously used to obtain classification models for natural/random proteins, breast/colon/prostate cancer-related proteins, DNA sequences of mycobacterial promoters and for early detection of diseases and drug-induced toxicities using the blood serum proteome mass spectrum. In order to show the extended practical potential of S2SNet, this work presents several examples of application for proteins, DNA/RNA, blood proteome mass spectra and time evolution of the financial law recurrence. The obtained topological indices can be used to characterize systems by creating classification models, clustering or pattern search with statistical, Neural Network or Machine Learning methods. The free availability of S2SNet, the flexibility of analyzing diverse systems and the Python portability make it an ideal tool in fields such as Bioinformatics, Proteomics, Genomics, and Biomedicine or Social, Economic and Political Sciences.

**Keywords:** Complex network, financial law network, graph indices, interaction, network, protein, python application, social network.

## 1. INTRODUCTION

The complexity of the real systems makes the comparison difficult between each other or the extraction of specific information that describes a discrete property. A possible strategy is to use the information about the connections or relationships between different parts of the entire system. This can be carried out using the Graph or Complex Network (CN) theory. A network is a collection of nodes represented by graphs composed of any items that have a) chemical, b) biological, c) social, and/or d) technological nature. For instance, atoms, molecules (proteins, DNA/RNA) are nodes of type a); viruses, bacteria, organisms, social actors and/or laws to regulate the behaviour of such actors may be considered as type b) and/or c); whereas computers, electric power plants, airports, mass spectra signals, or links between web pages or computers are usually allocated within the type d) of nodes [1-6]. This work presents four types of CN but we do not exclude other types of nodes and networks. In this approach we may use different invariant numbers often called Topological Indices (TIs) in order to encode or describe the structure of the systems. These TIs are usually derived from node-node adjacency or other types of matrices associated to the CN [7, 8]. Even though many TIs have been described only for molecular graphs of type a) many of them have been extended to be used in all types of CN [5, 9-13]. In any case, the development of new types of CN and graph representations or new TIs to describe them is an emerging field of science.

In the present work, a new tool for calculating TIs of a special type of CN is presented. S2SNet – Sequence to Star

*Address correspondence to this author at the Department of Information and Communications Technologies, Computer Science Faculty, University of A Coruña, Campus de Elviña s/n, 15071 A Coruña, Spain;
Tel: (+34) 981167000, Ext. 1302; Fax: (+34) 981167160;
E-mail: crm.publish@gmail.com

2212-392X/13 $58.00+.00 

Network [14] is a free Python desktop application for Microsoft Windows XP/Vista operating system. This new software has a friendly graphical user interface made with wXPython [15]. The software can be used to transform sequences of characters into a CN with Star Network (SN) topology. The SNs were introduced by Randic *et al.* in order to analyze the protein sequences, but have been used to investigate DNA/RNA or MS spectra of the blood proteome [16, 17] and may be extended to several other sequence-type data such as music, text, time series, etc. In this approach, the network/graph matrices are translated into the DOT language and plotted with Graphviz (*twopi*, *neato*, *dot*, *circo* and *fdp*) [18]. This paper introduces S2SNet as free software and gives detailed examples on how to use it in order to transform a text or numeric sequence into different types of SNs, visualize the resulted SNs, and calculate different classes of TIs for these SNs. The examples of sequences given here are the following: 1) the primary structure of the proteins, DNA or RNA and 2) the intensity of the Mass Spectra signals and 3) the time evolution of the recurrence to different types of laws in a financial law system. Thus, S2SNet can be used to describe systems in several fields such as social, economical, or political sciences, bioinformatics, structural biology, and clinical proteomics.

A previous application, March-Inside [19], has been used to calculate topological indices for Spiral and Star Graphs. However, that tool can generate Spiral Graphs from a list of sequences and Star Graphs only from the graphical interface, one by one. Other softwares such as Centibin [20] and Pajek [21] are calculating some of the S2SNet topological indices and process file by file in a matrix format (mat/net). Thus, S2SNet has the following advantages: transforms a list of sequences into Star Graph TIs at the click of a mouse, has the possibility to generate embedded graphs, calculates an extended list of Star Graph topological indices, and plots the resulted graphs in several Graphviz formats.

## 2. SOFTWARE DESCRIPTION

S2SNet is a free Python application that can transform character sequences into topological Star Network indices, transform number series into sequences, transform an N-character sequence into a 1-character sequence by changing the codification, edit/view the input/output txt files, create DOT language files, plot and display the networks as PNG images. Its basic architecture is described in Fig. (**1**).

S2SNet has two main panels, the principal window and the console output (see Fig. **2**). The main window has buttons for a fast access to the main features of the application:

- *S2SNet* - the transformation of sequences in SN TIs;

- *Help* - a short help page;

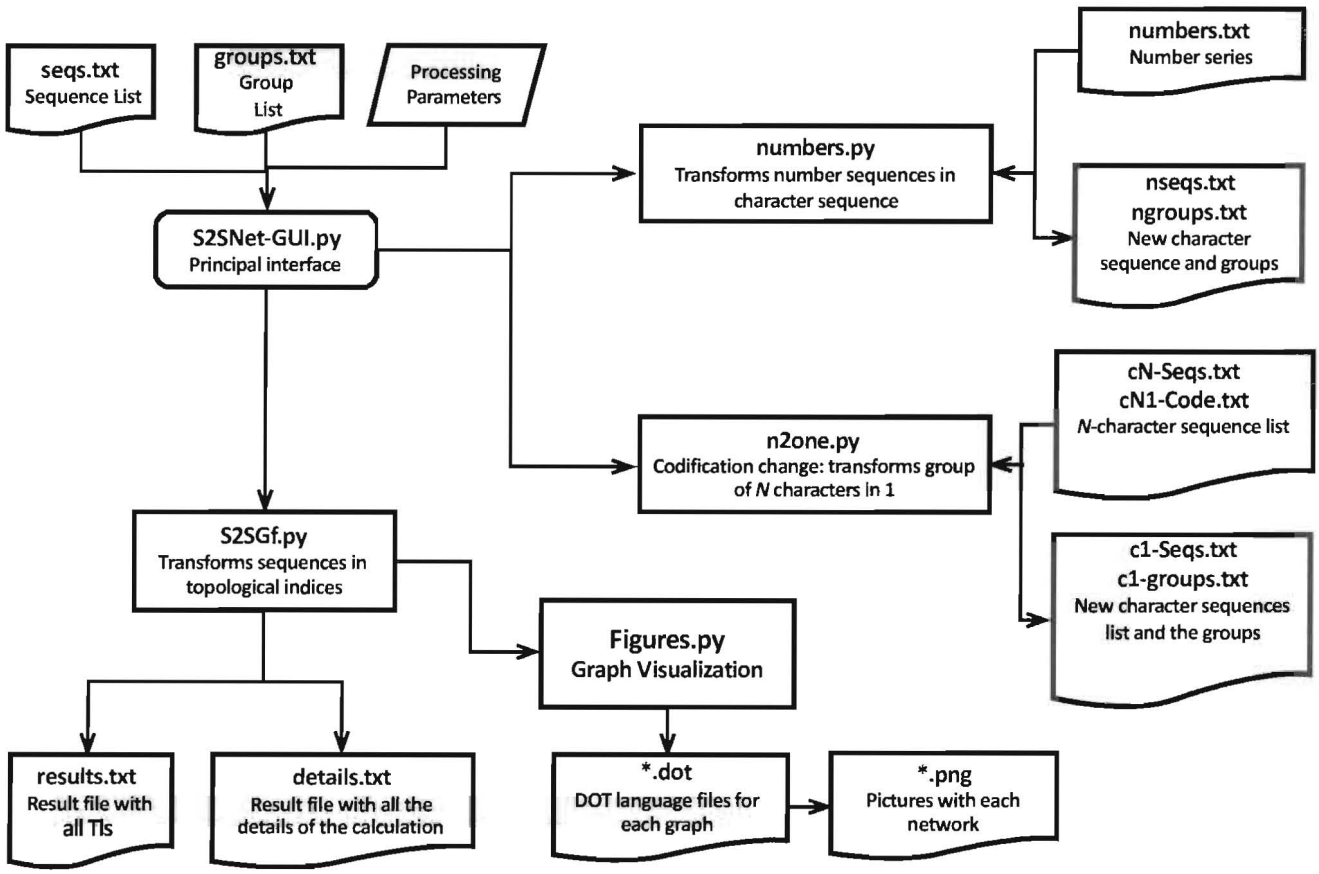



**Fig. (1).** Architecture of the S2SNet software: the tool has Python processing functions in, wxPython user interface and network representations created with the Graphviz package as graphical back-end.

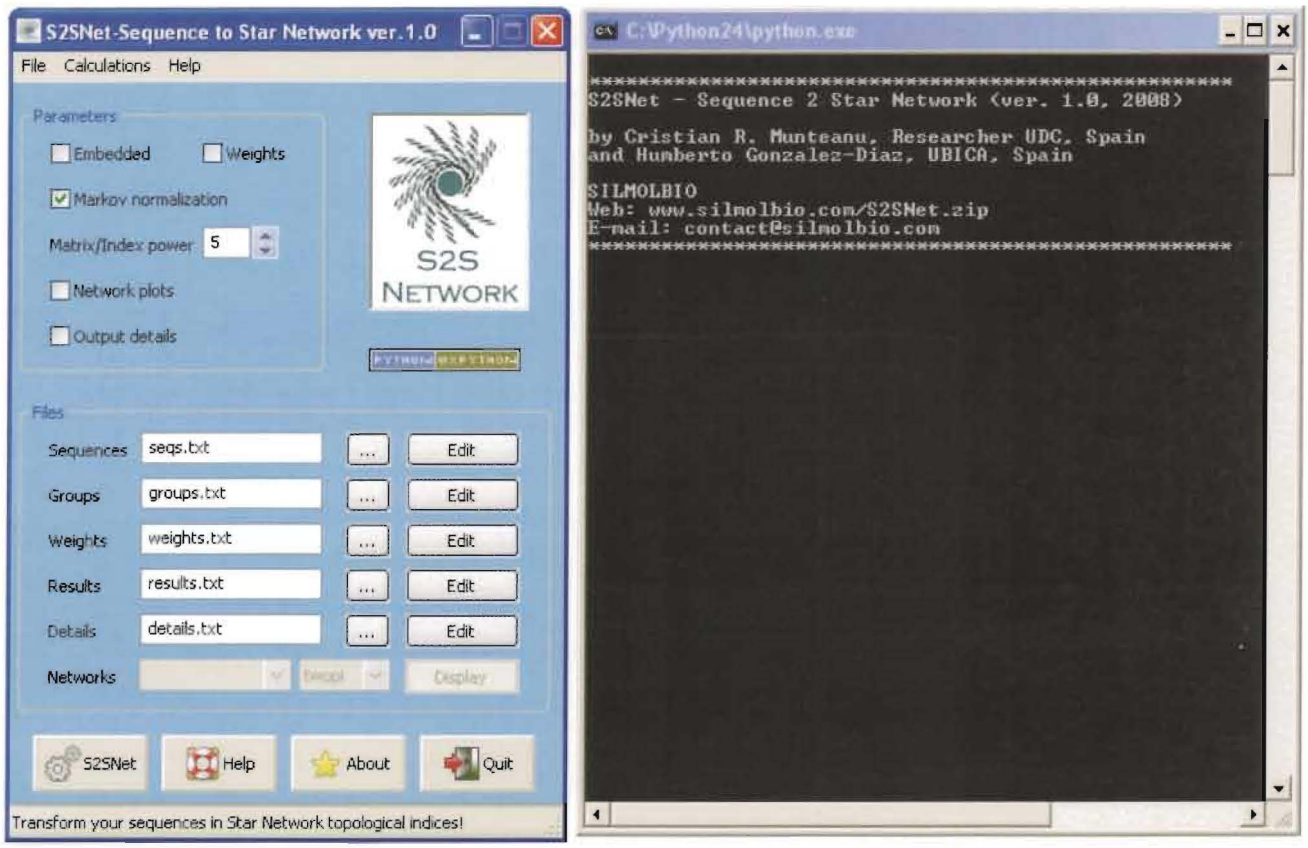



**Fig. (2).** S2SNet user interface.

- *About* - details about S2SNet and the authors;

- *Quit* - leave the tool.

The same options available in the S2SNet Menu divided into File, Calculations and Help.

In addition, the File and Calculations menus contain extra options:

- *New* – creates a new text document using the native Notepad from MS Windows;

- *Numbers to Sequence* – transforms the numerical input data such as the protein mass spectra into a 1-character sequence;

- *N to 1-Character Sequence* – changes the sequence codification by transforming N-character groups into 1-character sequences, such as the DNA/RNA codon sequences.

In the console panel, the details about the stage/errors/results of the calculations are displayed. In the main window you can choose the TIs calculation parameters, the input/output files and the visualization type of the resulted graphs:

- Parameters: embedded network, the use of the weight for each character; Markov normalization of the connectivity matrixes; if you want details of the calculations containing all the intermediate matrices and other info; power of the connectivity matrices and of some indices (max. 5); networks plotting support;

- Input files: sequences, groups and weights files;

- Output files: results and details files;

- Display mode for the Network plots: the sequence to display and the type of drawing application (*dot*, *circo*, *twopi*, *neato* and *fdp*); extra theoretical graphs are calculated and plotted such as the maximum and the average graphs of the sequences introduced as input.

All the options have default values for a calculation characterized by a non-embedded graph, a Markov normalizations of the matrices, a power of 5 for the matrices, no weight for the nodes, no network plot support and no detail file. In the displayed graphs, each group has a different color. If you need to obtain modified plots, you can find the DOT files (one for each sequence) and the Graphviz executables (*dot*, *circo*, *twopi*, *neato*, *fdp*) in the "dot" folder (if you enable the Network plot option).

The Calculation menu allows you to transform your data into the S2SNet format (1-character string):

- *Numbers to Sequence* (see Fig. **3**) - transforms your numbers into a sequence (the numbers must be TAB separated); you can choose the following parameters:

- The minimum and the maximum values of your data, number of groups you need (a maximum of 80); you have a GET button if you want to use the minimum and maximum calculated from your entire data;
- The input files: number (data) file;

- The output files: sequence files, group file and interval file (description of the group range).

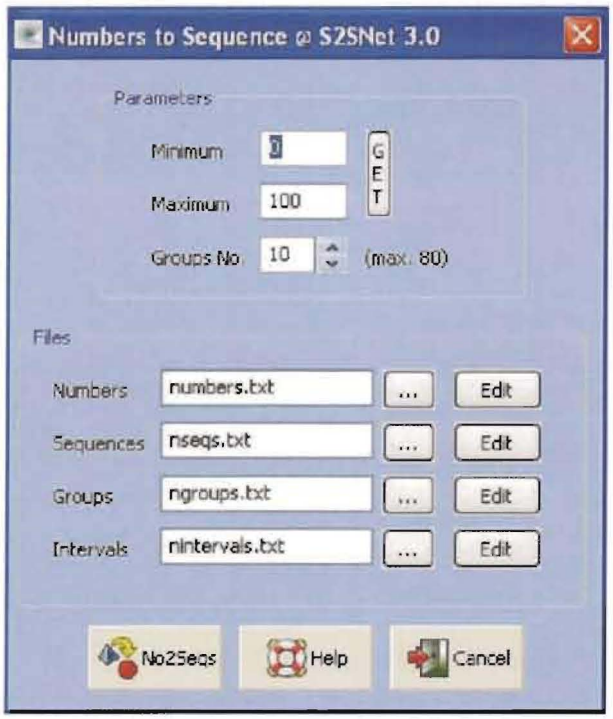



**Fig. (3).** Numbers to Sequence filter in S2SNet.

This filter can be used to transform the protein mass spectra numbers into text sequences and to calculate the corresponding Star Network indices.

- *N to 1-Character Sequence* (see Fig. **4**) - Transform your N-character sequences into 1-character sequences; you can set the following:

- The input files: N-character file (initial file), code file (the equivalence between N-character and 1-character; ex: ALA=A);
- The output files: 1-character files for S2SNet (final file) and group file (one item groups).

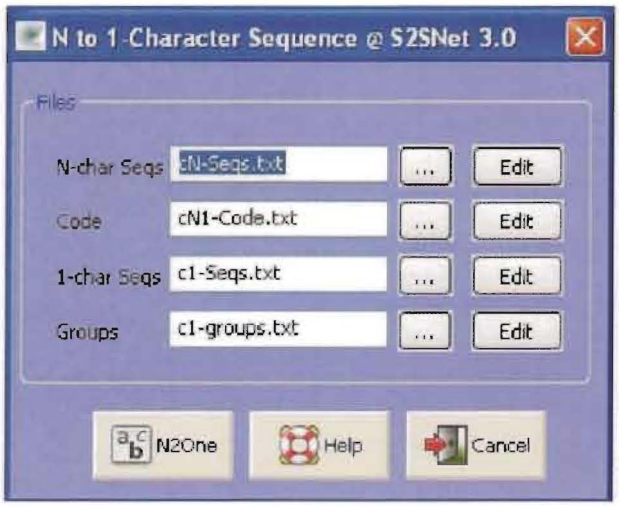



**Fig. (4).** N to 1-Character Sequence filter in S2SNet.

This filter can be used to transform sequences that contain items described by 3-letter codes such as the amino acids 3-letter code into sequences of 1-letter code; another example is the translation of the 3-nucleotides codons into the corresponding 1-letter amino acid sequence.

## 3. TIs OF SNs

In order to explain what an SN is, the protein representation is used in this section. Thus, a complex network can have the amino acids as vertices (nodes), connected in a specific sequence by the peptide bonds. The theoretical abstraction of a network is represented by a graph. The star graph is a special case of trees 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 [22]. In the case of proteins, each of the 20 possible branches ("rays") of the star contains the same amino acid type and the star center is a non-amino acid vertex. A protein primary sequence can be represented by different forms of graphs, which can be associated with distinct distance matrices (Randic *et al.*, 2007). The best method to construct a standard star graph is the following: each amino acid/vertex holds the position in the original sequence and the branches are labeled in the alphabetical order of the 3-letter amino acid code [17]. The embedded graph contains the initial sequence connectivity in the protein chain. Fig. (**5**) presents the non-embedded (A) and the embedded (B) star graphs of the HIV gp120 C5 protein (1meq: VKIEPLGVAPTKAKRRVVQREKR) using the alphabetical order of one-letter amino acid code. Thus, the primary structure of protein chains is transformed into the corresponding star graph invariant TIs. The resulted graphs do not depend on the 3-dimensional structure or the shape of the protein. The derived connectivity matrix, distance matrix and degree matrix are used to compare the graphs/networks. The matrices of the connectivity in the sequence and in the star graph are combined in the case of the embedded graph. These matrices and the normalized ones form the base for the TIs calculation.

The algorithm is begun by reading the sequences, groups and weights (see Fig. **1**). The transformation of each sequence into an SN is translated into a connectivity matrix ($M$), node degree vector ($deg$) and distance node matrix ($d$). If the graph is non-embedded $M$ it includes only the modified connectivity inside the SG. In the case of the embedded graphs, the original sequence connectivity will be added. The node degree represents the number of connections for a node in the graph and the distance matrices is filled with the number of nodes between each pair of nodes along the graph connections. If the weights option is chosen, $M$ will have the corresponding weight values along the matrix diagonal. In the Markov normalization (default option), $M$ is first normalized by dividing each element to the sum of the elements by row and after that raising it to the power $n$ given by the user (default is 5), resulting $n$ matrices ($M^n$). In case of non-Markov normalization, in the first step the matrix is powered, resulting $n$ matrices that are normalized by the same division.

These variables form the base for the calculation of the following TIs, presented in the output file (Todeschini and Consonni, 2002):

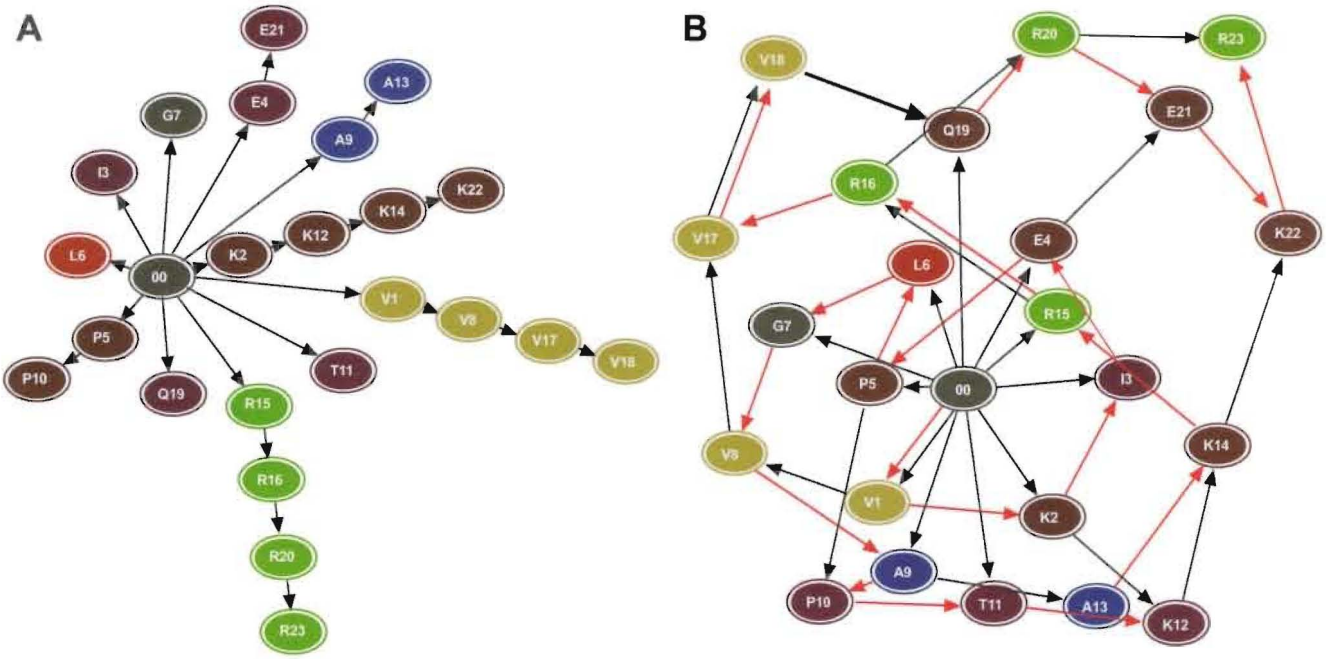



**Fig. (5).** The non-embedded (**A**) and embedded (**B**) graphs of HIV gp120 C5 protein.

- Shannon Entropy of the *n* Markov Matrices (*Sh*):

$$Sh_n = -\sum_i p_i * \log(p_i) \quad (1)$$

where $p_i$ are the $n_i$ elements of the vector $p$ resulting from the matrix multiplication of the powered Markov normalized matrix ($n_i$ x $n_i$) and a vector ($n_i$ x 1) with each element equal to $1/n_i$;

- Trace of the *n* connectivity matrices ($Tr_n$) or the spectral moments:

$$Tr_n = \sum_i (M^n)_{ii} \quad (2)$$

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

- Harary number (*H*) or the reciprocal distance sum index:

$$H = \sum_{i<j} m_{ij}/d_{ij} \quad (3)$$

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

- Wiener index (*W*) or the sum of the numbers of edges in the shortest paths in a graph between all pairs of amino acids in a protein:

$$W = \sum_{i<j} d_{ij} \quad (4)$$

- Gutman topological index ($S_6$):

$$S_6 = \sum_{ij} deg_i * deg_j \cdot d_{ij} \quad (5)$$

where $deg_i$ are the elements of the degree matrix;

- Schultz topological index (non-trivial part) (*S*):

$$S = \sum_{i<j} (deg_i + deg_j) * d_{ij} \quad (6)$$

- Balaban distance connectivity index (*J*) or average distance sum connectivity index (measures the graph ramification):

$$J = edges/(edges - nodes + 2) * \sum_{i<j} m_{ij} * sqrt(\sum_k d_{ik} * \sum_k d_{kj}) \quad (7)$$

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

- Kier-Hall connectivity indices ($^nX$):

$$^0X = \sum_i 1 / sqrt(deg_i) \quad (8)$$

$$^2X = \sum_{i<j<k} m_{ij} * m_{jk} / sqrt(deg_i * deg_j * deg_k) \quad (9)$$

$$^3X = \sum_{i<j<k<m} m_{ij} * m_{jk} * m_{km} / sqrt(deg_i * deg_j * deg_k * deg_m) \quad (10)$$

$$^4X = \sum_{i<j<k<m<o} m_{ij} * m_{jk} * m_{km} * m_{mo} / sqrt(deg_i * deg_j * deg_k * deg_m * deg_o) \quad (11)$$

$$^5X = \sum_{i<j<k<m<o<q} m_{ij} * m_{jk} * m_{km} * m_{mo} * m_{oq} / sqrt(deg_i * deg_j * deg_k * deg_m * deg_o * deg_q) \quad (12)$$

- Randic connectivity index ($^1X$):

$$^1X(R) = \sum_{ij} m_{ij} / sqrt(deg_i * deg_j) \quad (13)$$

These indices can be used in the next step in order to construct any classification model, clustering or pattern search using the statistical, neural networks or machine learning methods from applications such as STATISTICA [23] or WEKA [24].

## 4. RESULTS AND DISCUSSION

S2SNet has been successfully used by our group in four previous papers in order to create classification models for evaluating a protein as natural or random [25], as breast/colon [26] or prostate [27] cancer-related, a DNA sequence as a Mycobacterial promoter [28], and for early detection of diseases and drug-induced toxicities using the blood serum proteome mass spectrum [29]. This work presents four cases of S2SNet calculations for proteins, DNA, mass spectra and laws.

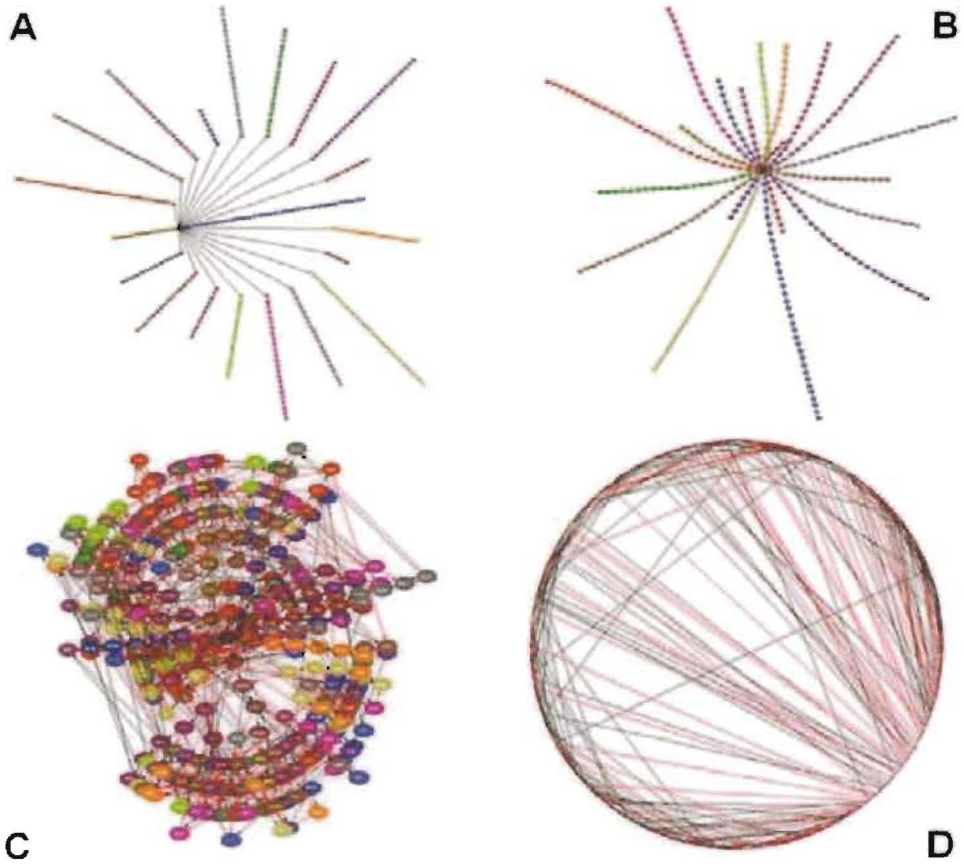



**Fig. (6).** 7ODCA star graphs: non-embedded SN created with twopi (**A**) and neato (**B**) and embedded SN created with twopi (**C**) and circo (**D**).

In the first case, the primary structure information of a protein sequence (the amino acid order and type) is encrypted in SN TIs. Fig. (**6**) shows the resulted graphs for 7ODC, chain A from the Protein Data Bank [30] obtained with four Graphwiz tools from S2SNet. The nodes represent the protein amino acids linked by the peptide bonds and are grouped into 20 branches corresponding to the different natural amino acids. The resulted TIs of the embedded SN are presented in Table **1**.

The next application is the DNA codons of COMT gene that is virtually translated with S2SNet into an amino acid sequence. This sequence is represented as an SN with the amino acids as nodes (equivalent to the codons) distributed in 21 branches, 20 standard amino acids and an extra X non-amino acid corresponding to the STOP DNA codons [31]. Thus, the primary structure of a DNA segment is transformed into SN TIs. The corresponding star network is presented in Fig. (**7A**). COMT (catechol-O-methyltransferase) is a gene that controls the function of the catechol-O-methyltransferase enzyme. This enzyme metabolizes catecholamines, which are heavily linked to dopaminergic and adrenergic/noradrenergic neurotransmission, or endorphins [32]. The variations of the COMT gene known as Single Nucleotide Polymorphism (SNP or SNIP) is thought by some researchers to be one of the key genes associated with schizophrenia [33].

Another example is the blood serum proteome mass spectrum that contains the patient's information about positive drug induced toxicity. They are then transformed into a sequence of characters (nodes), which define the group of SN upon the different range values of the signal intensity. The initial connectivity is generated by the signal positions in the spectrum. The embedded SNs of the COMT gene and proteome mass spectra are included in Fig. (**7**), and the corresponding embedded TIs are in Table **1**.

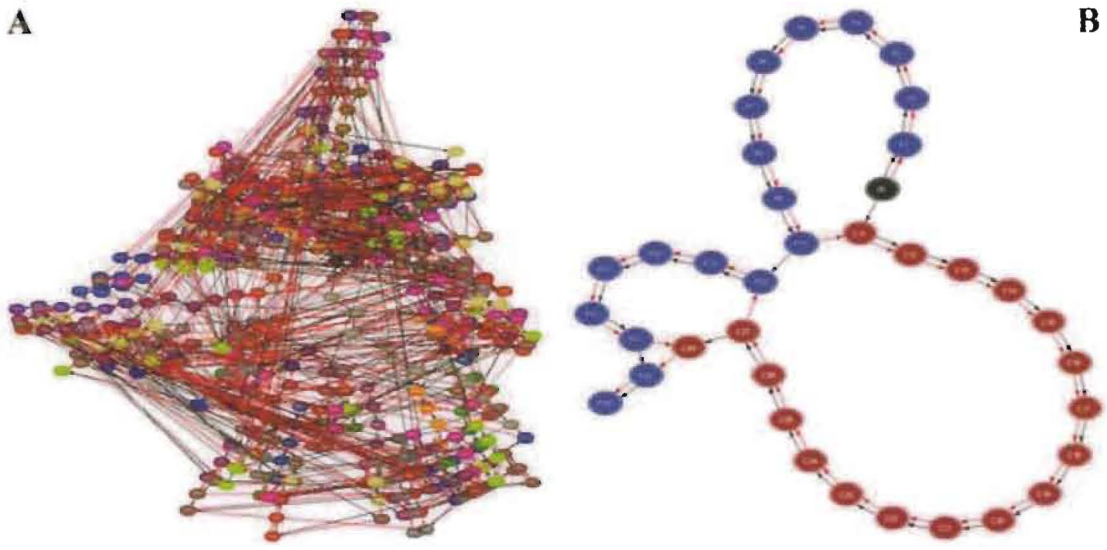
The last example of the S2SNet use is dedicated to the Social Network Analysis (SNA). SNA may be defined as the disciplined inquiry into patterning of relations among social actors, as well as the patterning of relationships among actors at different levels of analysis (such as persons and groups) [34]. It provides a common approach for all those disciplines involved in social structure study [35-38] susceptible of network depiction. Social structure concept is merely used in sociology and social theory. Although there is no agreement between theorists, it can refer to a specific type of relation between entities or groups, it can also evolve enduring patterns of behaviour and relationship within a society, or social institutions and regulations becoming embedded into social systems. For the most comprehensive review of SNA see the in-depth review of Newman M entitled *The Structure and Function of Complex Networks* [39]. In any case, considering that a network is a set of items,

**Table 1. The TIs of the Embedded SNs for the Protein 7ODC, the Gene COMT, and the Mass Spectrum of a Blood Proteome Sample of a Cancer Patient**

| TI | 7ODC Chain A Protein | COMT Gene | Blood Proteome Mass Spectrum |
|---|---|---|---|
| Sh0 | 5.94 | 6.63 | 3.59 |
| Sh1 | 6.20 | 6.82 | 3.68 |
| Sh2 | 6.15 | 6.79 | 3.67 |
| Sh3 | 6.21 | 6.84 | 3.69 |
| Sh4 | 6.19 | 6.83 | 3.68 |
| Sh5 | 6.21 | 6.85 | 3.69 |
| Tr0 | 388.00 | 769.00 | 37.00 |
| Tr1 | 0.00 | 0.00 | 0.00 |
| Tr2 | 101.03 | 202.65 | 17.22 |
| Tr3 | 1.48 | 6.19 | 0.22 |
| Tr4 | 46.84 | 95.93 | 12.27 |
| Tr5 | 2.30 | 7.77 | 0.35 |
| H | 2149.33 | 4785.45 | 118.46 |
| W | 9735114.00 | 75792640.00 | 8436.00 |
| S6 | 64058.51 | 140090.82 | 1113.23 |
| S | 75274220.00 | 578800473.00 | 36536.00 |
| J | 77495217.40 | 604518229.56 | 145898.28 |
| X0 | 199.40 | 396.63 | 25.55 |
| X1(R) | 192.71 | 382.74 | 18.36 |
| X2 | 184.62 | 367.67 | 13.04 |
| X3 | 176.73 | 353.05 | 9.12 |
| X4 | 168.77 | 338.93 | 6.27 |
| X5 | 160.91 | 325.34 | 4.31 |

usually called *nodes*, with connections between them, so-called *edges*, then we have a representation of social relationships in terms of nodes and ties, where nodes can be the individual actors within the networks, and ties the relationships between these actors [3]. In fact, SNA is nothing new in social sciences studies, as in the early 1930's, sociologists had already created a social network to study friendships between school children [40]. Since then, the importance of network approach to social sciences increased dramatically, and its applications expended from interrelation between family members [41] to company business interaction [42, 43] or patterns of sexual contacts [44, 45]. Although the network approach is so pervasive in the social sciences, its application in the law scope is still weak. Network tools and methodologies might be useful to illustrate the interrelation between different law types, and check the importance of a specific instrument so as the normative hierarchy respected by legislators. This helps to regulate the most important matter for individuals through law instruments which requires the approval from the most representative democratic actors. In this sense, the S2Snet software is a novel tool which has enabled herein the representation of the basic laws of the Spanish tax law system. Considering all this, we have built a graph on the recurrence to different types of laws related to tax matters over the years since 1946, see Fig. (**8**).

In these financial laws SN, we represented the time series for the use of different laws as a one-letter code sequence. In this sequence each specific type of rule is represented by a one letter code. Specifically, we transform all the regulation types as follows: Law (L), Instrument of Ratification (I), Decree-law (D), Order (O), Protocol (P), Royal Decree (R), Circular (C), EC Council Directive (E), Resolution (S), Organic Law (G), Agreement (A), Instruction (T), Convention (V), and Regulation (M). From there, a star-graph which connects several branches to a central node may be built. Each branch is composed of the same type of laws, and nodes -whether on the same branch or another being connected to each other if two laws are used at a time one




**Fig, (7).** Embedded SNs for the COMT gene (**A**) created with twopi and for the blood serum mass spectrum (**B**) carried out with neato.

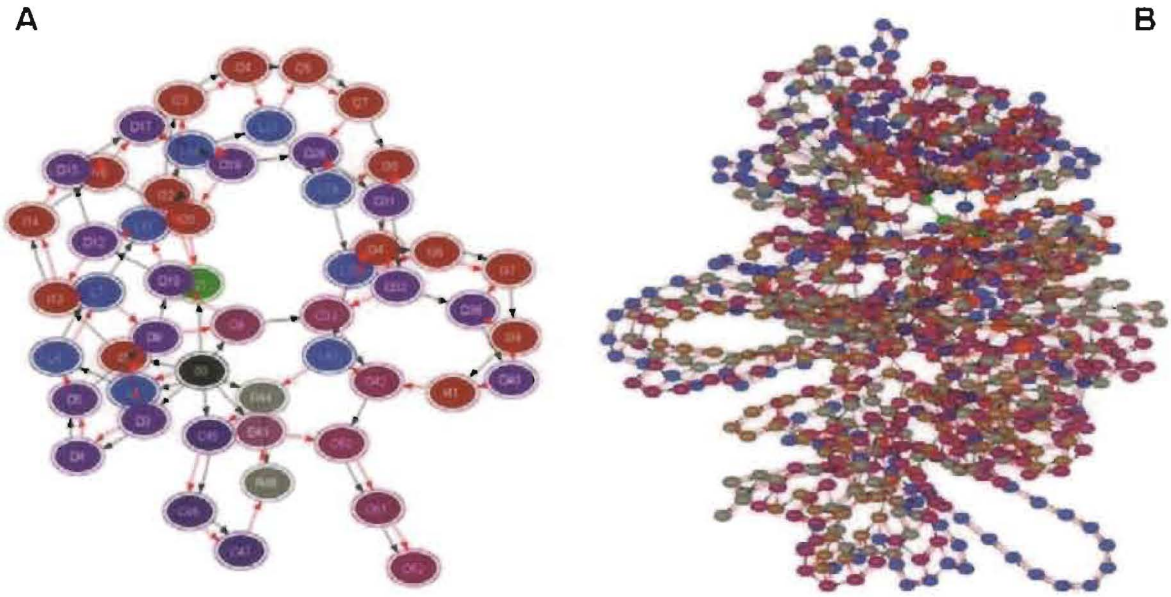



**Fig. (8).** Embedded SN for Spanish Financial law system over two periods: **A**) from 1946 to 1978 and **B**) from 1946 to 2004 (cumulative behavior with respect to 1946) using the neato algorithm.

after another. This method describes numerically the recurrence to different regulations or group of regulations or laws and can be used to describe the past dynamics and predict the future behaviour of the tax law application in Spain or other countries. In any case, many potential implications are still to be discovered in future research beyond this introductory work. In Table **2**, we illustrate the behaviour of different TIs for the SN constructed that reflect the changes in the recurrence to different laws in these periods.

## 5. CONCLUSIONS

This paper is proposing the use of S2SNet as a new tool in the studies of the SN calculating the TIs for characters or numbers sequences. It is a fast and free application that generates the invariant TI set for a specific sequence. These TIs can be used as the base for the development of Statistical, Neural Networks, or Machine Learning classification models/clustering/pattern search models. Thus, this tool can be used for e-learning or research in different fields such as Bioinformatics, Proteomics, Genomics, Biomedicine or Social, Economic, and Political Sciences. In this paper, we have presented only a few examples of possible SNs, but the application of S2SNet is not limited to any type of sequence.

## CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

## ACKNOWLEDGEMENTS

Cristian R. Munteanu acknowledges the funding support of the "Isidro Parga Pondal" Program from Xunta de Galicia (supported by the European Social Fund).

**Table 2. TIs of the Embedded SN for Spanish Financial Law System Over Two Periods: A) from 1946 to 1978 and B) from 1946 to 2004 (Cumulative Behavior with Respect to 1946), see also Fig. (8)**

| TI | 1946-1978 | 1946-2004 |
|---|---|---|
| Sh0 | 3.93 | 6.77 |
| Sh1 | 4.13 | 6.96 |
| Sh2 | 4.12 | 6.97 |
| Sh3 | 4.17 | 7.01 |
| Sh4 | 4.17 | 7.02 |
| Sh5 | 4.19 | 7.04 |
| Tr0 | 53.00 | 898.00 |
| Tr1 | 0.00 | 0.00 |
| Tr2 | 16.10 | 274.25 |
| Tr3 | 1.56 | 12.50 |
| Tr4 | 8.89 | 149.29 |
| Tr5 | 1.90 | 16.48 |
| H | 188.52 | 5727.39 |
| W | 24804.00 | 120691649.00 |
| S6 | 4380.84 | 130175.86 |
| S | 167069.00 | 812264716.00 |
| J | 192361.90 | 997661638.98 |
| X0 | 29.67 | 499.56 |
| X1(R) | 26.14 | 445.92 |
| X2 | 22.05 | 397.97 |
| X3 | 18.65 | 355.59 |
| X4 | 15.93 | 318.33 |
| X5 | 13.65 | 284.83 |

## REFERENCES


[1] Barabasi AL. Sociology. Network theory--the emergence of the creative enterprise. Science 2005; 308(5722): 639-41.

[2] Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004; 5(2): 101-13.

[3] Bornholdt S, Schuster HG. Handbook of Graphs and Complex Networks: From the Genome to the Internet. WILEY-VCH GmbH & CO. KGa: Wheinheim 2003.

[4] 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.

[5] Mason O, Verwoerd M. Graph theory and networks in Biology. IET Syst Biol 2007; 1(2): 89-119.

[6] Yook SH, Jeong H, Barabasi AL. Modeling the Internet's large-scale topology. Proc Natl Acad Sci USA 2002; 99(21): 13382-6.

[7] García-Domenech R, Gálvez J, Julián-Ortiz JV, Pogliani L. Some new trends in chemical graph theory. Chem Rev 2008; 108(3): 1127–69.

[8] Gonzalez-Diaz H, Gonzalez-Diaz Y, Santana L, Ubeira FM, Uriarte E. Proteomics, networks and connectivity indices. Proteomics 2008; 8(4): 750-78.

[9] Concu R, Podda G, Uriarte E, González-Díaz H. A new computational chemistry & complex networks approach to structure - function and similarity relationships in protein enzymes. In: Handbook of Computational Chemistry Research; Collett CTaR, CD., ed., Nova Science Publishers 2009.

[10] Dall'asta L, Alvarez-Hamelin I, Barrat A, Vazquez A, Vespignani A. Statistical theory of Internet exploration. Phys Rev E Stat Nonlin Soft Matter Phys 2005; 71(3 Pt 2A): 036135.

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

[12] 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.

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

[14] Munteanu CR, Gonzáles-Diáz H. S2SNet - Sequence to Star Network, Reg. No. 03 / 2008 / 1338, Santiago de Compostela, Spain. Santiago de Compostela, Spain 2008.

[15] Rappin N, Dunn R. wxPython in Action. Manning Publications Co.: Greenwich, CT 2006.

[16] Ferino G, Delogu G, Podda G, Uriarte E, González-Díaz H. Quantitative Proteome-Disease Relationships (QPDRs) in Clinical Chemistry: Prediction of Prostate Cancer with Spectral Moments of PSA/MS Star Networks. In: Clinical Chemistry Research; Mitchem BHaS, Ch.L., ed., Nova Science Publisher: NY 2009.

[17] Randić M, Zupan J, Vikic-Topic D. On representation of proteins by star-like graphs. J Mol Graph Model 2007: 290-305.

[18] Koutsofios E, North SC. Drawing Graphs with dot. AT&T Bell Laboratories, Murray Hill: NJ, USA 1993.

[19] González-Díaz H, Molina-Ruiz R, Hernandez I. March-Inside version 3.0 (Markov Chains Invariants for Simulation & Design); Windows supported version under request to the main author contact email: gonzalezdiazh@yahoo.es. 3.0 ed 2007.

[20] Koschützki D. CentiBiN Version 1.4.2. 2006: CentiBiN Version 1.4.2, Centralities in Biological Networks © 2004-6 Dirk Koschützki Research Group Network Analysis, IPK Gatersleben, Germany.

[21] Batagelj V, Pajek M.A. Program for Large Network Analysis (ver. 1.15), http://vlado.fmf.uni-lj.si/pub/networks/pajek/. 1.15 ed 2006.

[22] Harary F. Graph Theory. Westview Press: MA 1969.

[23] StatSoft.Inc. STATISTICA, (data analysis software system), version 6.0, www.statsoft.com. 6.0 ed 2002.

[24] Witten IH, Frank E. WEKA: Waikato Environment for Knowledge Analysis. 2000.

[25] Munteanu CR, Gonzalez-Diaz H, Borges F, de Magalhaes AL. Natural/random protein classification models based on star network topological indices. J Theor Biol 2008; 254(4): 775-83.

[26] Munteanu CR, Magalhaes AL, Uriarte E, Gonzalez-Diaz H. Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. J Theor Biol 2009; 257(2): 303-11.

[27] González-Díaz H, Ferino G, Munteanu CR, Vilar S, Uriarte E. Protein Graphs in Cancer Prediction. In: Oncoproteomics; Cho WK, ed., Springer 2009.

[28] Perez-Bello A, Munteanu CR, Ubeira FM, De Magalhaes AL, Uriarte E, Gonzalez-Diaz H. Alignment-free prediction of mycobacterial DNA promoters based on pseudo-folding lattice network or star-graph topological indices. J Theor Biol 2009; 256(3): 458-66.

[29] Cruz-Monteagudo M, Munteanu CR, Borges F, *et al.* Stochastic molecular descriptors for polymers. 4. Study of complex mixtures with topological indices of mass spectra spiral and star networks: The blood proteome case. Polymer 2008; 49: 5575-87.

[30] Berman HM, Westbrook J, Feng Z, *et al.* The Protein Data Bank. Nucleic Acids Res 2000; 28: 235-42.

[31] Griffiths AJF, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM. Introduction to Genetic Analysis. 7th ed. W. H. Freeman & Co.: New York 1999.

[32] Zubieta JK, Heitzeg MM, Smith YR, *et al.* COMT val158met genotype affects mu-opioid neurotransmitter responses to a pain stressor. Science 2003; 299(5610): 1240-3.

[33] Wonodi I, Mitchell BD, Stine OC, *et al.* Lack of association between COMT gene and deficit/nondeficit schizophrenia. Behav Brain Funct 2006; 2: 42.

[34] Breiger R. The Analysis of Social Networks. In: Handbook of Data Analysis; Hardy M, Bryman A, eds., Sage Publications: London 2004; 505-26.

[35] Abercrombie N, Hill S, Turner BS. The Penguin Dictionary of Sociology. In: Social structure, Penguin: London 2000.

[36] Craig C. Social Structure, Dictionary of the Social Sciences,. Oxford University Press: Oxford 2002.

[37] Wellman B, Berkowitz SD. Social Structures: A Network Approach. Cambridge University Press: Cambridge 1988.

[38] White H, Boorman S, Breiger R. Social Structure from Multiple Networks: Blockmodels of Roles and Positions. American Journal of Sociology 1976; 81(730-780).

[39] Newman M. The structure and function of complex networks. SIAM Review 2003; 56: 167-256.

[40] Moreno JL. Who Shall Survive?. Beacon House: New York 1934.

[41] Padgett JF, Ansell CK. Robust Action and the Rise of the Medici, 1400-1434. Amer J Sociol 1993; 98(6): 1259-319.

[42] Mariolis P. Interlocking directorates and control of corporations: The theory of bank control. Social Sci Quart 1975; 56: 425-39.

[43] Mizruchi MS. The American Corporate Network, 1904-1974. Sage: Beverly Hills 1982.

[44] Klovdahl AS, Potterat JJ, Woodhouse DE, Muth JB, Muth SQ, Darrow WW. Social networks and infectious disease: the Colorado Springs Study. Soc Sci Med 1994; 38(1): 79-88.

[45] Liljeros F, Edling CR, Amaral LA, Stanley HE, Aberg Y. The web of human sexual contacts. Nature 2001; 411(6840): 907-8.


Received: July 23, 2011 Revised: September 15, 2011 Accepted: September 16, 2011

**Artículo 2.**

**Predicción de Redes Complejas con Índices Markov-Wiener y ANNs**

Modeling Complex Metabolic Reactions, Ecological Systems, and Financial and Legal Networks with MIANN Models Based on Markov-W iener Node Descriptors. A. Du ardo-Sánchez, C.R. Munteanu, P. Riera-Fernández, A. López-Díaz, A. Pazos, and H. González-Díaz. *Journal of Chemical Information and Modelling*, **2014**, 54, 16-29

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

# Modeling Complex Metabolic Reactions, Ecological Systems, and Financial and Legal Networks with MIANN Models Based on Markov-Wiener Node Descriptors

Aliuska Duardo-Sánchez,[†,‡] Cristian R. Munteanu,*[,†] Pablo Riera-Fernández,[†] Antonio López-Díaz,[‡] Alejandro Pazos,[†] and Humberto González-Díaz[§,∥]

[†]Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruña, Campus de Elviña, 15071, A Coruña, A Coruña, Spain

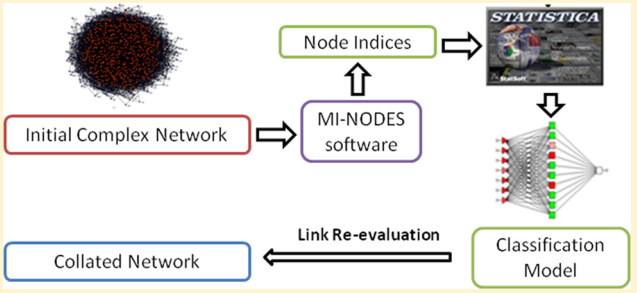
[‡]Department of Special Public Law, Financial and Tributary Law Area, Faculty of Law, University of Santiago de Compostela (USC), 15782, Santiago de Compostela, A Coruña, Spain

[§]Department of Organic Chemistry II, Faculty of Science and Technology, University of the Basque Country (UPV/EHU), 48940, Leioa, Bizkaia, Spain

[∥]IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Biscay, Spain

Supporting Information

**ABSTRACT:** The use of numerical parameters in Complex Network analysis is expanding to new fields of application. At a molecular level, we can use them to describe the molecular structure of chemical entities, protein interactions, or metabolic networks. However, the applications are not restricted to the world of molecules and can be extended to the study of macroscopic nonliving systems, organisms, or even legal or social networks. On the other hand, the development of the field of Artificial Intelligence has led to the formulation of computational algorithms whose design is based on the structure and functioning of networks of biological neurons. These algorithms, called Artificial Neural Networks (ANNs), can be useful for the study of complex networks, since the numerical parameters that encode information of the network (for example centralities/node descriptors) can be used as inputs for the ANNs. The Wiener index ($W$) is a graph invariant widely used in chemoinformatics to quantify the molecular structure of drugs and to study complex networks. In this work, we explore for the first time the possibility of using Markov chains to calculate analogues of node distance numbers/$W$ to describe complex networks from the point of view of their nodes. These parameters are called Markov-Wiener node descriptors of order $k^{th}$ ($W_k$). Please, note that these descriptors are not related to Markov-Wiener stochastic processes. Here, we calculated the $W_k(i)$ values for a very high number of nodes (>100,000) in more than 100 different complex networks using the software MI-NODES. These networks were grouped according to the field of application. Molecular networks include the Metabolic Reaction Networks (MRNs) of 40 different organisms. In addition, we analyzed other biological and legal and social networks. These include the Interaction Web Database Biological Networks (IWDBNs), with 75 food webs or ecological systems and the Spanish Financial Law Network (SFLN). The calculated $W_k(i)$ values were used as inputs for different ANNs in order to discriminate correct node connectivity patterns from incorrect random patterns. The MIANN models obtained present good values of Sensitivity/Specificity (%): MRNs (78/78), IWDBNs (90/88), and SFLN (86/84). These preliminary results are very promising from the point of view of a first exploratory study and suggest that the use of these models could be extended to the high-throughput re-evaluation of connectivity in known complex networks (collation).




## 1. INTRODUCTION

**1.1. The Classic Wiener Index.** In the last part of the nineteenth century and in the twentieth century the interest for the study of the molecular structure led to the formulation of questions about how to encode and quantify the information contained in the molecule. As a result of these questions, the concept of molecular descriptor was defined as "the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment".[1] In 1947, Wiener published an article entitled *Structural determination of paraffin boiling points*.[2] In this work it is proposed that organic compounds, as well as all their physical properties, depend functionally upon the number, kind, and structural arrangement of the atoms in the molecule.

**Received:** May 10, 2013
**Published:** December 8, 2013





dx.doi.org/10.1021/ci400280n | *J. Chem. Inf. Model.* 2014, 54, 16−29

Therefore, it is possible to find an equation that relates the structure of the studied paraffins with their boiling points. This equation, also used by Wiener in refs 3−5, can be written in a general form as

$$\Delta t = \frac{a}{n^2}\cdot \Delta w + b\cdot \Delta p \quad (1)$$

where $n$ is the number of atoms, $p$ is the polarity number, defined as the number of pairs of carbon atoms which are separated by three carbon bonds, and $w$ is the path number, defined as the sum of the distances between any two carbon atoms in the molecule and considered as one of the oldest topological indices. This last term was coined by Hosoya in 1971 to refer to the Z index,[6,7] and it is currently used to define all the numerical quantifiers of molecular topology that are mathematically derived from the structural graph of a molecule, usually an H-depleted molecular graph.[8] The path number is also called Wiener index or Wiener number ($W$), and it is calculated as the half sum of all the elements $d_{ij}$ of the distance matrix ($D$). As it can be seen, more distant atom pairs make a larger contribution to $W$ than adjacent atom pairs:

$$W = \frac{1}{2}\cdot \sum_{i=1}^{D}\sum_{j=1}^{D} d_{ij} \quad (2)$$

It is interesting to point out that the Wiener index was independently proposed in 1959 by Harary in the context of sociometry, with the name *total status of a graph*[9] as well as in 1975 by Rouvray and Crafford.[10] This fact indicates that the Wiener index concept was not well-known in those times. However, in the middle of the 1970s many authors began to study the properties and applications of molecular descriptors. This led to the development of new topological indices (TIs), some of which were based on $W$. A complete list and a brief explanation of these indices can be found in ref 11. Taking into consideration the classification proposed by Balaban for the TIs,[12] $W$ (together with Z) belongs to the first of three generations or classes. First-generation TIs are integer numbers based on integer local graph-vertex invariants (LOVI) and have a high degeneracy that limits their use.

According to Diudea and Gutman,[10] the physical and chemical properties of organic substances, which can be expected to depend on the area of the molecular surface and/or on the branching of the molecular carbon-atom skeleton, are usually well correlated with $W$. Among them, there are the heats of formation, vaporization and atomization, density, boiling point, critical pressure, refractive index, surface tension and viscosity of various, acyclic and cyclic, saturated and unsaturated as well as aromatic hydrocarbon species, velocity of ultrasound in alkanes and alcohols, rate of electroreduction of chlorobenzenes, etc. Correlations between $W$ and melting points were also reported, but in this case the results were not completely satisfactory. $W$ was also used to predict the behavior of organic substances in gas chromatography, for instance, chromatographic retention times (CRT) of monoalkyl- and *o*-dialkylbenzenes. In this sense, $W$ is very useful in chemoinformatics for the search of models that connect the molecular structure with molecular properties.

**1.2. Applications of the Wiener Index in Chemoinformatics.** In pharmaceutical design, we can find many applications of the Wiener index. For instance, Mandloi et al.[13] investigated the correlation of Wiener ($W$), Szeged (Sz), and molecular connectivity indices ($^0\chi_R$, $^1\chi_R$, and $^2\chi_R$) with molecular properties. Log P values of benzoic acid and its nuclear-substituted derivatives were used for this purpose. The statistical analyses for univariate and multivariate correlations indicated that both $W$ and Sz are closely related to the connectivity indices ($^m\chi_R$) and that $W$, Sz, and $^1\chi_R$ have similar modeling potentials ($^1\chi_R$ gives slightly better results than both $W$ and Sz). $^0\chi_R$ and $^2\chi_R$ are poorly correlated with log P. Lukovits established correlations between $W$ and cytostatic and antihistaminic activities of certain pharmacologically interesting compounds as well as between $W$ and their Estron-binding affinities.[14] He also employed $W$ in the study of the *n*-octanol/water partition coefficient.[15]

Mendiratta and Madan[16] studied the relationship between $W$ and the antiviral activity of a series of 118 5-vinylpyrimidine nucleoside analogues. The predicted activity of each compound was compared with reported antiviral activity against herpes simplex virus type I. Due to the significant correlation between antiviral activity and $W$, it was possible to predict antiviral activity with an accuracy of 83%.

In the work carried out by Agrawal et al.[17] the antimalarial activity of a series of sulfonamide derivatives (2,4-diamino-6-quinazoline sulfonamides) was modeled topologically using $W$ and Sz. It was observed that the models based on $W$ gave slightly better results than the models based on Sz. Sardana and Madan[18] studied the relationship of the molecular connectivity index ($^1\chi$), $W$, and the eccentric connectivity index ($\xi^c$) with the diuretic activity of 68 sulfamoylbenzoic acid derivatives. The models had an 82% accuracy rate in $^1\chi$, an 85% accuracy rate in $W$, and a 90% accuracy rate in $\xi^c$. In another work, the relationship of $W$, Zagreb group parameter (M1), and $\xi^c$ with the anticonvulsant activity of a series of 41 substituted benazamides/benzylamines was investigated.[19] The models had an 88% (M1), 94% ($\xi^c$), and 97% ($W$) accuracy rate.

Gupta et al. studied[20] the relationship of $\xi^c$ and $W$ with regard to anti-inflammatory activity for a data set consisting of 76 pyrazole carboxylic acid hydrazide analogues. A prediction with a 90% accuracy rate was obtained using $\xi^c$ and an 84% accuracy rate in the case of $W$. Bajaj et al.[21] studied the relationship of the Wiener topochemical index (a modification of $W$ sensitive to the presence of heteroatoms and with less degeneracy) and Wiener index with the anti-HIV activity of 62 phenethylthiazolethiourea compounds. The prediction accuracy rate was 90% in both cases. The relationship of anti-HIV activity of 61 acylthiocarbamates with $W$, $^1\chi$, and $\xi^c$ was also investigated by these authors.[22] 95% ($\xi^c$), 97% ($^1\chi$), and 98% ($W$) accuracy rates were observed. In another work, the relationship of anti-inflammatory activity of 112 N-arylanthranilic acids with $W$, Zagreb indices M1 and M2, and $\xi^c$ was studied.[23] The different models had an 82.6% ($\xi^c$), 86.8% ($W$), 88.88% (M1), and 90.3% (M2) accuracy rate

In the area of cancer research, the inhibition of CDK2/cyclin A by 42 3-aminopyrazoles was studied using $W$, the atomic molecular connectivity index ($\chi^A$), and the superadjacency topochemical index ($\int^{Ac}$).[24] The different models had an 86% ($W$), 88% ($\int^{Ac}$), and 89% ($\chi^A$) accuracy rate. With the aim to develop methods to select drug candidates for the treatment of Alzheimer's disease, Kumar and Madan studied[25] the relationship of $W$, M1, and $\xi^c$ with the glycogen synthase kinase-3 beta inhibitory activity of 28 thiadiazolidinones. The prediction accuracy rate was 83% (M1), 86% ($\xi^c$), and 87% ($W$). Finally, Lather and Madan studied[26] the relationship between $W$ and multidrug-resistance-associated protein inhibitory activity of 82 pyrrolopyrimidines and their derivatives. The prediction

accuracy rate of the model was 88%. As we can see in these examples, the Wiener index has a wide range of applications in predictive studies, and, since it is one of the first topological indices, it is used in many works in order to compare the performance of new introduced indices.

**1.3. Complex Networks and MARCH-INSIDE Models.** Graph and Complex Network theory is expanding its application to different levels of matter organization such as molecular, biological, technological, and social networks.[27−29] A network is a set of items, usually called *nodes*, with connections between them, which are called *links* or *edges*.[30] The nodes can be atoms, molecules, proteins, nucleic acids, drugs, cells, organisms, parasites, people, words, laws, computers, or any other part of a real system. The edges or links are relationships between the nodes, such as chemical bonds, physical interactions, metabolic pathways, pharmacological actions, law recurrence, or social ties.[31−39]

On the other hand, there are many different experimental and/or theoretical methods to assign node−node links depending on the type of network we want to construct. Unfortunately, many of these methods are expensive in terms of time or resources. In addition, different methods that link nodes in the same type of network are not totally accurate and consequently they do not always coincide. A possible solution to this problem is the use of node descriptors of known networks as inputs of predictive models.[40] The reasons for using re-evaluations of link connectivity in networks are the following:

1. The experimental networks can have errors due to experiment conditions, calibrations, human errors, etc.

2. There are networks where the connectivity is just a prediction or it is the result of text data mining techniques (all involve possible errors).

3. The model that can re-evaluate the node connectivity can be used for new nodes as an alternative to the expensive and time-consuming experiments. In some cases, such as the interaction of all the possible pairs − triples of molecules, it is impossible to be carried out experimentally.

4. Contradictory information for nodes and links for different networks.

In fact, the use of predictive models in which the inputs are graph parameters is not limited to the study of molecules and has been extended to other complex systems.[41,42] The first and one of the most studied TIs is the Wiener index, and it is possible to use Markov Chains (MC) to calculate it locally or globally within a graph considering all possible branches at different topological distances. The information is quantified in terms of $W_k(j)$ values, which are called Markov-Wiener node descriptors of order $k^{th}$ for all $j^{th}$ states (nodes) of an MC associated with the system. This MC is expressed by a Markov or Stochastic matrix ($\mathbf{\Pi}_1$) and represented by a graph of the studied system. The elements of $\mathbf{\Pi}_1$ are the probabilities ${}^1p_{ij}$ with which the $i^{th}$ and $j^{th}$ nodes connect to each other (there is a physical or functional tie, link, or relationship) within a graph. By using Chapman-Kolmogorov equations it is straightforward to realize the way to calculate $W_k(j)$ values for all nodes in a graph. We can use these values directly or sum some of them to obtain total or local parameters. Our group has introduced the software called MARCH-INSIDE (Markovian Chemicals In Silico Design), which has become a very useful tool for predictive studies on drugs, proteins, and more complex systems.[43−57] This software can calculate 1D (sequence), 2D (connectivity in the plane), and 3D (connectivity in the space) MC parameters, including $W_k(j)$ values, for many molecular systems. MARCH-INSIDE is able to characterize small molecules (drugs, metabolites, organic compounds), biopolymers (gene sequence, protein sequence or 3D structure, and RNA secondary structure) and artificial polymers, but it can perform a limited manage of other complex networks. This occurs because MARCH-INSIDE can read, transform into Markov matrix, represent as graph, and calculate the Wiener index for molecular formats (.mol or SMILE .txt files for drugs, .pdb for proteins, or .ct files for RNAs), but it is unable to upload formats of Complex Networks (.mat, .net, .dat, .gml, etc.).

Consequently, we have reprogrammed the MARCH-INSIDE application, creating new software able to manage complex networks. The new program is called MI-NODES (MARCH-INSIDE NOde DEScriptors), and it is compatible with other programs like Pajek or CentiBin, since it is able to read .mat, .net, and .dat formats. A very interesting feature of MI-NODES is that it can process multiple networks and calculate both MC global TIs and/or node descriptors for all these networks. It is also able to export them in a single file in network-by-network and/or node vs node output formats. The classic TIs can include additional information such as Markov node linkage probability ($p_{ij}$) for any $i$, $j$ nodes of a graph. In previous works, we have introduced other types of Markov TIs (and the respective node descriptors): Markov-Shannon Entropy node descriptors,[58] Markov-Randić indices,[59] Markov-Rücker indices,[60] Markov-Galvez indices,[61] Markov-Autocorrelation node descriptors,[62] and Markov-Harary numbers.[63] In these previous studies, we have used the indices, calculated with MI-NODES, in order to compare several types of complex networks from different fields such as biology, linguistics, technology, sociology, and law.

**1.4. MIANN Models.** The methods used to predict structure−property relationships in complex systems (molecular or not) can be classified into two types: methods of type (1), used to quantify the structure of the system and methods of type (2), able to link the structure of the system with a property of this system (and others). Several methods of type (1) use Quantum Mechanics (in Molecular Sciences) and/or Graph theory (in Molecular and Social Sciences as well), whereas the methods of type (2) use Statistical and/or Machine Learning (ML) techniques.[64−68] Many computer programs implement type (1) and/or type (2) methods with applications in Molecular Sciences and/or a wide range of areas depending on the flexibility of the algorithms used. For instance, DRAGON,[69−71] TOPS-MODE,[72−75] TOMOCOMD,[76,77] CODESSA,[78,79] and MOE[80] are classic programs used to apply type (1) methods in Molecular Sciences. CentiBin,[81] Pajek,[82] or MI-NODES implement type (1) methods with applications in almost all areas of sciences but at the cost of simplification of detailed representation of the system. On the other hand, the Linear Discriminant Analysis (LDA) implemented in STATISTICA[83] or the ML methods implemented in WEKA[84] are examples of type (2) methods with widespread applications. In this context, different researchers/journals have edited important monographic issues in order to discuss different computational methods. For instance, Bisson has edited a special issue about Computational Chemogenomics in drug design and discovery.[85] Speck-Planche and Cordeiro guest-edited a special issue about computer-aided, synthesis and assay of anticancer agents.[86] Prado-Prado and García-Mera have also guest-edited a special issue about computer-aided drug design and molecular docking for disorders of the central nervous system and other diseases.[87] González-Díaz has guest-edited two special issues

about multitarget models and Complex Networks applied to medicinal chemistry.[88,89] In all these issues, and others of the same journal, several review and research papers in this area[52,80,90−123] have been published.

In particular, the bioinspired Artificial-Intelligence (AI) algorithms called Artificial Neural Networks (ANNs) are among the most powerful type (2) methods. As we mentioned in the previous section, MARCH-INSIDE (MI from now on) is a well-known type (1) method mentioned in many recent works published by different groups.[39,52,124−130] We can combine MI with different Machine Learning algorithms. In particular, we can combine MI with ANNs in order to seek predictive models. The name of this strategy is MIANN (MI and ANN models). In a recent paper, we have reviewed the MIANN strategy including theoretical basis, implementation in Web servers, and examples of applications in molecular sciences.[131] We have also developed new MIANN models for drug-target interactions, several physicochemical properties of surfactants, and large reaction networks in organic synthesis.

In the present work we introduce for the first time a new type of Wiener-like indices called Wiener-Markov node descriptors. This algorithm of type (1) is implemented in MI-NODES. Then, we use for the first time the MIANN strategy to study complex biomolecular, ecological, and social and legal systems using the Wiener-Markov node descriptors as input. In order to illustrate the use of the new method we have carried out three studies. In each study, we report for the first time a new model useful to re-evaluate the connectivity quality of different types of networks. Although very different systems were studied, the same workflow was used in all the experiments (see Figure 1).

The idea is to search for a MIANN model that uses the $W_k(j)$ values calculated for the nodes of a complex network as inputs of ANNs to decide which nodes are correctly linked. This class of model will allow us to computationally re-evaluate all the links of nodes in any complex network so that we do not have to rely upon experimentation to confirm the existence or not of a link between all pairs of links. By using this model, we should experimentally confirm only those connections predicted by the model with low link score and/or simply remove them from the network depending on the cost/benefit ratio. This work is proposing three studies: each study is proposing a prediction model based on several networks of the same network type. In the first study, we processed the full set of metabolic reactions of different organisms (bacteria, yeast, nematode, and plants). The node descriptors from 40 networks represented the model data set. In the second study, we used different biological and ecological networks, including predator−prey, parasite-host, plant-seed disperser, anemone-clown fish species, and others. In the last study, we illustrate the application of the method to a complex network that takes into account all the historical record (1940−2004) of the Spanish Financial Law system (legal and social network). With the advent of the age of complex system sciences, this work can be considered as a basis for a relatively little studied but very important field: the assessment of the connectivity quality in new complex networks.

## 2. MATERIALS AND METHODS

**2.1. Markov-Wiener Node Descriptors.** The classic Markov matrix ($^1\mathbf{\Pi}$) for each network is constructed as follows: first, we download from the Internet the connectivity matrix **L** or the data about the links between the nodes to assemble **L** ($n \times n$ matrix, where $n$ is the number of vertices).

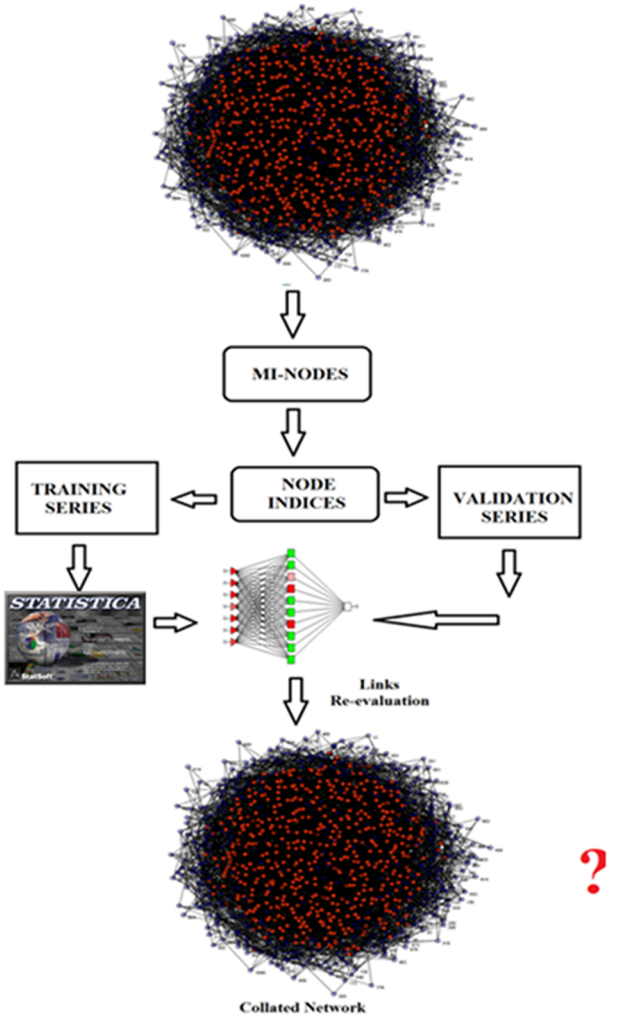



**Figure 1.** MIANN workflow example: blue/red nodes are training/validation cases (dark/light in gray scale).

Next, the Markov matrix $\mathbf{\Pi}$ is built. It contains the vertices probability ($p_{ij}$) based on **L**. The probability matrix is raised to the power $k$, resulting $(^1\mathbf{\Pi})^k$. The resulting matrices $^k\mathbf{\Pi}$ are the $k^{\text{th}}$ natural powers of $^1\mathbf{\Pi}$ and contain the transition probabilities $^kp_{ij}$. These are the probabilities to reach the $j^{\text{th}}$ node moving from the $i^{\text{th}}$ node throughout a walk of length $k$ (for each $k$). The generalization of the classic $W$ to general Markov-Wiener indices of order $k^{\text{th}}$ is straightforward to carry out by multiplying the values of $d_{ij}$ (distances obtained from the distance matrix $D$) by these probabilities $^kp_{ij}$. Therefore, we can obtain $k$ values of the new Markov-Wiener indices $W_k(G)$ for a graph $G$, instead of only one Wiener index value obtained with the classic formulation. In addition, we can run the sum only over all the $j^{\text{th}}$ nodes linked to one specific node $i$ (the number of these nodes is symbolized here as $j \rightarrow i$ and it is equal to $\delta_i$, the degree of $i$). In this simple case we can obtain a total of $k$ values of new Markov-Wiener node descriptors, $W_k(i)$, for the node $i^{\text{th}}$:

$$W_k(G) = \frac{1}{2}\cdot\sum_{i=1}^{\Pi D}\sum_{j=i}^{\Pi D} {}^kp_{ij}\cdot d_{ij} \tag{3}$$

$$W_k(i) = \frac{1}{2}\cdot\sum_{i=1}^{1}\sum_{j\rightarrow i}^{\delta_i} {}^kp_{ij}\cdot d_{ij} = \frac{1}{2}\cdot\sum_{j\rightarrow i}^{\delta_i} {}^kp_{ij}\cdot d_{ij} \tag{4}$$


dx.doi.org/10.1021/ci400280n | J. Chem. Inf. Model. 2014, 54, 16−29

Table 1. Average Values $W_k(i)_{org.avg}$ of Metabolic Networks of 43 Organisms vs Classic Parameters[a]

| organism | $N$ | $L_{in}$ | $L_{out}$ | $R$ | $E$ | $g_{in}$ | $g_{out}$ | $D$ | $W_1$ | $W_2$ | $W_3$ | $W_4$ | $W_5$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aae | 419 | 1278 | 1249 | 401 | 285 | 2.1 | 2.2 | 3.3 | 0.87 | 1.08 | 1.26 | 1.44 | 1.57 |
| Aac | 395 | 1202 | 1166 | 380 | 271 | 2.1 | 2.2 | 3.2 | 0.88 | 1.07 | 1.24 | 1.43 | 1.56 |
| Afu | 496 | 1527 | 1484 | 486 | 299 | 2.2 | 2.2 | 3.5 | 0.85 | 1.09 | 1.29 | 1.48 | 1.61 |
| Ape | 204 | 588 | 575 | 178 | 135 | 2.2 | 2.2 | 3.2 | 0.95 | 1.11 | 1.25 | 1.46 | 1.6 |
| Ath | 302 | 804 | 789 | 250 | 185 | 2.1 | 2.3 | 3.5 | 0.89 | 1.12 | 1.3 | 1.48 | 1.62 |
| Bbu | 187 | 442 | 438 | 140 | 106 | 2.3 | 2.4 | 3 | 0.8 | 0.99 | 1.18 | 1.37 | 1.49 |
| Bsu | 785 | 2794 | 2741 | 916 | 516 | 2.2 | 2.1 | 3.3 | 0.8 | 1.09 | 1.3 | 1.52 | 1.65 |
| Cac | 494 | 1624 | 1578 | 511 | 344 | 2.1 | 2.2 | 3.3 | 0.83 | 1.08 | 1.28 | 1.46 | 1.59 |
| Cel | 462 | 1446 | 1418 | 450 | 295 | 2.1 | 2.2 | 3.3 | 0.9 | 1.12 | 1.32 | 1.51 | 1.65 |
| Cje | 380 | 1142 | 1115 | 359 | 254 | 2.1 | 2.3 | 3.2 | 0.88 | 1.09 | 1.27 | 1.45 | 1.58 |
| Cte | 389 | 1097 | 1062 | 333 | 231 | 2.1 | 2.2 | 3.3 | 0.88 | 1.1 | 1.3 | 1.51 | 1.63 |
| Cpn | 194 | 401 | 391 | 134 | 84 | 2.2 | 2.3 | 3.4 | 0.99 | 1.14 | 1.27 | 1.47 | 1.62 |
| Ctr | 215 | 479 | 462 | 158 | 94 | 2.2 | 2.4 | 3.5 | 0.9 | 1.06 | 1.22 | 1.38 | 1.5 |
| Csp | 546 | 1782 | 1746 | 570 | 370 | 2 | 2.2 | 3.3 | 0.88 | 1.13 | 1.33 | 1.56 | 1.68 |
| Dra | 815 | 2870 | 2811 | 965 | 557 | 2.2 | 2.1 | 3.3 | 0.89 | 1.12 | 1.31 | 1.52 | 1.65 |
| Eco | 778 | 2904 | 2859 | 968 | 570 | 2.2 | 2.1 | 3.2 | 0.79 | 1.03 | 1.24 | 1.44 | 1.57 |
| Efa | 386 | 1244 | 1218 | 382 | 281 | 2.1 | 2.2 | 3.1 | 0.81 | 1.04 | 1.24 | 1.42 | 1.55 |
| Eni | 383 | 1095 | 1081 | 339 | 254 | 2.1 | 2.2 | 3.3 | 0.89 | 1.11 | 1.31 | 1.5 | 1.65 |
| Hin | 526 | 1773 | 1746 | 597 | 361 | 2.1 | 2.3 | 3.2 | 0.77 | 1.05 | 1.26 | 1.48 | 1.59 |
| Hpy | 375 | 1181 | 1144 | 375 | 246 | 2 | 2.3 | 3.3 | 0.89 | 1.11 | 1.3 | 1.5 | 1.62 |
| Mbo | 429 | 1247 | 1221 | 391 | 282 | 2.2 | 2.2 | 3.2 | 0.87 | 1.09 | 1.27 | 1.46 | 1.6 |
| Mge | 209 | 535 | 525 | 196 | 85 | 2.4 | 2.2 | 3.5 | 0.96 | 1.14 | 1.26 | 1.38 | 1.48 |
| Mja | 424 | 1317 | 1272 | 415 | 264 | 2.2 | 2.3 | 3.5 | 0.88 | 1.11 | 1.29 | 1.47 | 1.6 |
| Mle | 422 | 1271 | 1244 | 402 | 282 | 2.2 | 2.2 | 3.2 | 0.83 | 1.06 | 1.25 | 1.44 | 1.58 |
| Mpn | 178 | 470 | 466 | 154 | 88 | 2.3 | 2.2 | 3.2 | 0.91 | 1.11 | 1.29 | 1.46 | 1.59 |
| Mtu | 587 | 1862 | 1823 | 589 | 358 | 2 | 2.2 | 3.3 | 0.88 | 1.12 | 1.32 | 1.55 | 1.67 |
| Ngo | 406 | 1298 | 1270 | 413 | 285 | 2.1 | 2.2 | 3.2 | 0.85 | 1.06 | 1.24 | 1.42 | 1.56 |
| Nme | 381 | 1212 | 1181 | 380 | 271 | 2.2 | 2.2 | 3.2 | 0.86 | 1.08 | 1.27 | 1.45 | 1.59 |
| Osa | 292 | 763 | 751 | 238 | 178 | 2.1 | 2.3 | 3.5 | 0.93 | 1.19 | 1.39 | 1.57 | 1.71 |
| Pae | 734 | 2453 | 2398 | 799 | 490 | 2.1 | 2.2 | 3.3 | 0.87 | 1.1 | 1.29 | 1.52 | 1.65 |
| Pfu | 316 | 901 | 867 | 283 | 191 | 2 | 2.3 | 3.4 | 0.93 | 1.14 | 1.33 | 1.5 | 1.65 |
| Pgi | 424 | 1192 | 1156 | 374 | 254 | 2.2 | 2.2 | 3.3 | 0.85 | 1.06 | 1.24 | 1.41 | 1.54 |
| Pho | 323 | 914 | 882 | 288 | 196 | 2 | 2.2 | 3.4 | 0.92 | 1.12 | 1.31 | 1.49 | 1.63 |
| Spn | 416 | 1331 | 1298 | 412 | 288 | 2.1 | 2.2 | 3.2 | 0.86 | 1.08 | 1.25 | 1.44 | 1.57 |
| Rca | 670 | 2174 | 2122 | 711 | 427 | 2.1 | 2.2 | 3.4 | 0.92 | 1.12 | 1.27 | 1.5 | 1.63 |
| Rpr | 214 | 510 | 504 | 155 | 100 | 2.3 | 2.3 | 3.4 | 0.91 | 1.11 | 1.27 | 1.44 | 1.57 |
| Sce | 561 | 1934 | 1889 | 596 | 402 | 2 | 2.2 | 3.3 | 0.88 | 1.11 | 1.31 | 1.54 | 1.68 |
| Spy | 403 | 1300 | 1277 | 404 | 280 | 2.1 | 2.2 | 3.1 | 0.89 | 1.08 | 1.24 | 1.44 | 1.57 |
| Mth | 430 | 1374 | 1331 | 428 | 280 | 2.2 | 2.2 | 3.4 | 0.89 | 1.13 | 1.33 | 1.52 | 1.65 |
| Tma | 338 | 1004 | 976 | 302 | 223 | 2.1 | 2.2 | 3.2 | 0.88 | 1.09 | 1.28 | 1.47 | 1.6 |
| Tpa | 207 | 562 | 555 | 175 | 124 | 2.2 | 2.3 | 3.1 | 0.86 | 1.03 | 1.21 | 1.42 | 1.55 |
| Sty | 819 | 3008 | 2951 | 1007 | 577 | 2.2 | 2.2 | 3.2 | 0.82 | 1.06 | 1.26 | 1.46 | 1.59 |
| Ype | 568 | 1754 | 1715 | 580 | 386 | 2.1 | 2.2 | 3.3 | 0.86 | 1.08 | 1.26 | 1.45 | 1.59 |

[a]Note: $N$ = number of substrate, $L$ = number of links, $R$ = number of individual reactions or temporary substrate-enzyme complexes, $E$ = number of enzymes, $g_{in}$ and $g_{out}$ = the exponents, $D$ = diameter of the metabolic network.

**2.2. Data sets Used.** *2.2.1. Metabolic Reaction Networks (MRNs).* The data were downloaded directly from Barabasi's group Web site (http://www.nd.edu/~networks/resources.htm) as gzipped ASCII files. In these files each number represents a substrate in the metabolic network. Data-format is as follows: From → To (directed link). The information studied was previously obtained by Jeong et al. from the 'intermediate metabolism and bioenergetics' portions of the WIT database and used in order to try to understand the large-scale organization of metabolic networks.[132] According to the authors, the biochemical reactions described within the WIT database are composed of substrates and enzymes connected by directed links. For each reaction, educts and products were considered as nodes connected to the temporary educt−educt complexes and associated enzymes. Bidirectional reactions were considered separately. For a given organism with N substrates, E enzymes, and R intermediate complexes the full stoichiometric interactions were compiled into an (N+E+R) X (N+E+R) matrix, generated separately for each of the different organisms. Table 1 shows a summary of the properties of the MRNs studied. The names, abbreviations, and links for all the networks studied are as follows: *Aquifex aeolicus* = Aae; *Actinobacillus actinomycetemcomitans* = Aac; *Archaeoglobus fulgidus* = Afu; *Aeropyrum pernix* = Ape; *Arabidopsis thaliana* = Ath; *Borrelia burgdorferi* = Bbu; *Bacillus subtilis* = Bsu; *Clostridium acetobutylicum* = Cac; *Caenorhabditis elegans* = Cel; *Campylobacter jejuni* = Cje; *Chlorobium tepidum* = Cte; *Chlamydia pneumoniae* = Cpn; *Chlamydia trachomatis* = Ctr;


dx.doi.org/10.1021/ci400280n | J. Chem. Inf. Model. 2014, 54, 16−29

*Synechocystis sp.* = Csp; *Deinococcus radiodurans* = Dra; *Escherichia coli* = Eco; *Enterococcus faecalis* = Efa; *Emericella nidulans* = Eni; *Haemophilus influenza* = Hin; *Helicobacter pylori* = Hpy; *Mycobacterium bovis* = Mbo; *Mycoplasma genitalium* = Mge; *Methanococcus jannaschii* = Mja; *Mycobacterium leprae* = Mle; *Mycoplasma pneumonia* = Mpn; *Mycobacterium tuberculosis* = Mtu; *Neisseria gonorrheae* = Ngo; *Neisseria meningitidis* = Nme; *Oryza sativa* = Osa; *Pseudomonas aeruginosa* = Pae; *Pyrococcus furiosus* = Pfu; *Porphyromonas gingivalis* = Pgi; *Pyrococcus horikoshii* = Pho; *Streptococcus pneumonia* = Spn; *Rhodobacter capsulatus* = Rca; *Rickettsia prowazekii* = Rpr; *Saccharomyces cerevisiae* = Sce; *Streptococcus pyogenes* = Spy; *Methanobacterium thermoautotrophicum* = Mth; *Thermotoga maritima* = Tma; *Treponema pallidum* = Tpa; *Salmonella typhi* = Sty; *Yersinia pestis* = Ype.

*2.2.2. Interaction Web Database Biological Networks (IWDBNs).* The IWDB (http://www.nceas.ucsb.edu/interactionweb/resources.html) contains data sets on species interactions from several communities in different parts of the world. In a recent review, we have discussed and listed many biological networks including those contained in the IWDB.[63] Data include many types of ecological interactions: plant-pollinator, plant-frugivore, plant-herbivore, plant-ant mutualism, and predator−prey interactions. Most webs are "bipartite networks", which consist of two groups that are assumed to interact with species in the other group but not with species within their own group (e.g., plants and insect herbivores). Almost all data sets or webs (ecological network) are presented with an "interaction matrix" format (type 1 matrices), in which columns represent one group (e.g., plants) and rows represent the other group (e.g., pollinators). The exceptions to this format are predator−prey (food) webs, which are "one-mode" webs, represented by a symmetric matrix with all species listed in both columns and rows (type 2 matrices). In a previous work, we downloaded and transformed all matrices into .net format, which list all pairs (arcs or edges) of species (nodes) into a text file.[133] Later, we uploaded all .net files of all ecological networks to calculate numerical parameters using the MI-NODES software. This tool processes all the .net files as matrices (see the next sections). Table 2 shows a summary of all the available data sets with reference to their sources.

*2.2.3. Spanish Financial Law Network (SFLN).* The studied network is built establishing connections between two laws or legal regulations (nodes) if the time-lag is less than 1 for the same type of laws. Consequently, law-law links represent the corecurrence of different regulations in the Spanish Financial System over time, which depend in turn on social and economical conditions. The Spanish financial law recurrence network associated with the matrix **L** with elements $L_{ij}$ was reported in previous works.[134]

**2.3. MI-NODES Software for the Calculation of Markov-Wiener Node Descriptors.** MI-NODES (MARCH-INSIDE NOde DEScriptors) is a GUI Python/wxPython application used for the calculation of node descriptors/topological indices of nodes, subnetworks, or full networks. Actually, it should be considered as the generalization of the MARCH-INSIDE software adapted to manage any kind of complex networks (this program was originally designed to study drugs, proteins, and nucleic acid structures). MI-NODES calculates new types of node descriptors ${}^{k}C_{c}(j)$ based on Markov normalized node probabilities without a prior removal of each node to perform calculations. It also calculates Markov generalizations of different topological indices ${}^{k}TI_{c}(G)$ of class $c$ and power $k$ for the graph $G$. The tool is both Pajek and CentiBin compatible, since it is able to read networks in the following formats: .net, .dat, and .mat.

**2.4. MIANN Models.** Let $S_j$ be the output variable of a model used to score the quality of the connectivity pattern $L_{ij}$ between the node $i^{th}$ and all the remnant $(n-1)$ nodes in the network. In this sense, $S_j$ is a real valued variable that scores the quality of the connectivity pattern or links (all direct and indirect connections) established between the node $j^{th}$ and the other nodes. The higher is the value of $S_j$ the closer to the correct pattern are the links set for $j^{th}$ in the network as a whole, according to the model. On the other hand, $L_j$ is the input dependent variable. $L_j = 1$ when a node is correctly linked to the rest of the nodes in the network, and $L_j = 0$ when a node has a random connectivity model. We can use ANNs to search for a nonlinear and/or linear equation with coefficients $a_k$, ${}^{g}b_k$, and $c_0$. In the particular case of a linear equation, obtained by means of a Linear Neural Network (LNN),[135] the general formula can be written as

$$S_j = \sum_{k=0}^{5} a_k \cdot W_k(j) + \sum_{g=0}^{g=Ng} \sum_{k=0}^{5} b_{gk} \cdot [W_k(j) - W_k(j)_{g.avg}] + c_0$$
$$= \sum_{k=0}^{5} a_k \cdot W_k(j) + \sum_{g=0}^{g=Ng} \sum_{k=0}^{5} b_{gk} \cdot \Delta W_k(j)_g + c_0 \quad (5)$$

In this equation we can see the coefficients $(a_k)$ of the Wiener-Markov node descriptors used as input $W_k(j)$ and/or the coefficients $({}^{g}b_k)$ of different deviation terms constructed with these variables. The deviation terms have the general form $\Delta W_k(j)_g = [W_k(j) - W_k(j)_{g.avg}]$, where $W_k(j)_{g.avg}$ is the average value (avg) of $W_k(j)$ for a subset or group $(g)$ of nodes of the same graph $G$ $(g \in G)$ that obey a given condition. This type of deviation terms resembles the moving average terms used in time series models like in Box−Jenkins' ARIMA models.[136] However, in the present work $g$ may be not only a time frame or season (laws approved in the same year) but also a biological boundary (metabolic reactions in the same organism) or spatial condition (interactions in the same eco-system); see the Results section.

The linear equation of the MIANN model obtained by means of LNN for MRNs is

$$S_j = \sum_{k=0}^{5} a_k \cdot W_k(j) + \sum_{k=0}^{5} b_{gk} \cdot [W_k(j) - W_k(j)_{Org.avg}] + c_0$$
$$= \sum_{k=0}^{5} a_k \cdot W_k(j) + \sum_{g=0}^{g=Ng} \sum_{k=0}^{5} b_{gk} \cdot \Delta W_k(j)_{Org} + c_0 \quad (6)$$

The LNN model for the particular case of IWDBNs has the following formula:

$$S_j = \sum_{k=0}^{5} a_k \cdot W_k(j) + \sum_{k=0}^{5} b_{gk} \cdot [W_k(j) - W_k(j)_{Web.avg}] + c_0$$
$$= \sum_{k=0}^{5} a_k \cdot W_k(j) + \sum_{k=0}^{5} b_{gk} \cdot \Delta W_k(j)_{Web} + c_0 \quad (7)$$

Finally, the LNN model for the particular case of SFLN has the following formula


dx.doi.org/10.1021/ci400280n | J. Chem. Inf. Model. 2014, 54, 16−29

**Table 2. Summary of Almost All Data Sets Included in the IWDB**

| data set[a] | habitat type | location | data type[b] | #OA[c] | #OB[c] |
|---|---|---|---|---|---|
| Anemone − Fish Networks | | | | | |
| 1 | coral reefs | Indo-Pacific | binary | 10 | 26 |
| Host − Parasite Networks | | | | | |
| 2 | freshwater lake | Canada | pii | 7 | 29 |
| 3 | freshwater lake | Canada | pii | 10 | 40 |
| 4 | freshwater lake | Canada | prevalence | 31 | 144 |
| 5 | river | Canada | pii | 14 | 51 |
| 6 | river | Canada | pii | 17 | 53 |
| 7 | freshwater lake | Canada | prevalence | 33 | 97 |
| 8 | freshwater reservoir | Canada | pii | 6 | 25 |
| 45 | salt marsh | USA | binary | | |
| Plant − Ant Networks | | | | | |
| 9 | rainforest | Australia | no. visits | 51 | 41 |
| 10 | rainforest | Peru | no. visits | 8 | 18 |
| 11 | tropical forest | Costa Rica | no. visits | 6 | 4 |
| 12 | Amazon rainforest | Brazil | no. visits | 16 | 25 |
| Plant[a] − Herbivore Networks | | | | | |
| 13 | arid grasslands | USA | binary | 54 | 24 |
| | | | | 52 | 22 |
| 14 | whole country | Finland | binary | 5 | 64 |
| | | Britain | | 6 | 88 |
| Plant − Pollinator Networks | | | | | |
| 15 | Andean scrub | Chile | binary | 87 | 98 |
| | | | | 43 | 62 |
| | | | | 41 | 28 |
| 16 | boreal forest | Canada | I. caught | 12 | 102 |
| 17 | caatinga[b] | Brazil | no. visits | 13 | 13 |
| 18 | montane forest and grass land | USA | binary | 96 | 276 |
| 19 | high-altitude desert | Canary Islands | binary | 11 | 38 |
| 20 | Alpine subarctic community | Sweden | no. visits | 23 | 118 |
| 21 | Arctic community | Canada | binary | 29 | 86 |
| 22 | heathland habitat heavily invaded by introduced plants | Mauritius Island | rates | 135 | 74 |
| | heathland habitat with plants removed | | | 100 | 64 |
| 23 | beech forest | Japan | I. caught | 93 | 679 |
| 24 | high Arctic | Canada | no. visits | 32 | 115 |
| 25 | montane forest | Australia | I. caught | 42 | 91 |
| 26 | multiple communities | Galapagos Islands | binary | 106 | 54 |
| 27 | xeric scrub | Argentina | binary | 21 | 45 |
| | woody riverine vegetation and xeric scrub | | | 23 | 72 |
| 28 | meadow | United Kingdom | F. of visits | 25 | 79 |
| 29 | Arctic community | Canada | I. caught | 11 | 18 |
| Plant − Pollinator Networks | | | | | |
| 30 | deciduous forest | USA | no. visits | 13 | 44 |
| 31 | coastal forest | Mauritius Island | no. visits | 14 | 13 |
| | rocky cliff and open herb community | Azores Islands | | 10 | 12 |
| 32 | upland grassland | South Africa | I. caught | 9 | 56 |
| 33 | palm swamp community | Venezuela | binary | 33 | 53 |
| 34 | agricultural area | USA | binary | 456 | 1429 |
| 35 | caatinga | Brazil | binary | 51 | 25 |
| 36 | maple-oak woodland | USA | no. visits | 7 | 32 |
| 37 | peat bog | Canada | I. caught | 13 | 34 |
| 38 | evergreen montane forest | Argentina | no. visits | 10 | 29 |
| | | | | 9 | 33 |
| | | | | 9 | 27 |
| | | | | 10 | 29 |
| | | | | 8 | 35 |
| | | | | 8 | 26 |
| | | | | 7 | 24 |
| | | | | 8 | 27 |
| Plant − Seed Disperser Networks | | | | | |
| 39 | forest | Papua New Guinea | no. visits | 31 | 9 |
| 40 | semideciduous tropical forest | Panama | F. removed | 13 | 11 |
| 41 | primary montane tropical rainforest | Kenya | no. visits | 19 | 71 |
| | secondary montante tropical rainforest | | | 14 | 57 |
| | canopy | | | 15 | 71 |
| | midstory | | | 8 | 34 |
| | understory | | | 10 | 33 |
| 42 | neotropical forest | Trinidad | no. visits | 65 | 14 |
| 43 | - | - | no. visits | 22 | 20 |
| 44 | temperate woodland | United Kingdom | no. visits | 12 | 14 |
| Predator − Prey Food Webs | | | | | |
| 45 | salt marsh | USA | | binary | 128 |
| 46 | pine forest | New Zealand | | binary | 85 |
| | pasture grassland | | | | 87 |
| | | | | | 95 |
| | | | | | 109 |
| | tussock grassland | | | | 107 |
| | broadleaf forest | | | | 78 |
| | | | | | 78 |
| | pine forest | USA | | | 78 |
| | | | | | 105 |
| | | | | | 71 |
| | | | | | 58 |

[a]Data set name: 1 = Ollerton et al. (2007); 2 = Aishihik Lake; 3 = Cold Lake; 4 = Lake of the Woods; 5 = McGregor River; 6 = Parsnip River; 7 = Lake Huron; 8 = Smallwood Reservoir; 9 = Blüthgen et al. (2004); 10 = Davidson et al. (1989); 11 = Davidson and Fisher (1991); 12 = Fonseca and Ganade (1996); 13 = Joern (1979); 14 = Leather (1991); 15 = Arroyo et al. (1982); 16 = Barrett and Helenurm (1987); 17 = Bezerra et al. (2009); 18 = Clements and Long (1923); 19 = Dupont et al. (2003); 20 = Elberling and Olesen (1999); 21 = Hocking (1968); 22 = Kaiser-Bunbury et al. (2009); 23 = Kato et al. (1990); 24 = Kevan (1970); 25 = Inouye and Pyke (1988); 26 = McMullen (1993); 27 = Medan et al. (2002); 28 = Memmott (1999); 29 = Mosquin and Martin (1967); 30 = Motten (1982); 31 = Olesen et al. (2002); 32 = Ollerton et al. (2003); 33 = Ramírez and Brito (1992); 34 = Robertson (1929); 35 = Santos et al. (2010); 36 = Schemske et al. (1978); 37 = Small (1976); 38 = Vázquez and Simberloff (2002); 39 = Beehler (1983); 40 = Poulin et al. (1999); 41 = Schleuning et al. (2010); 42 = Snow and Snow (1971); 43 = Snow and Snow (1988); 44 = Sorensen (1981); 45 = Lafferty et al. (2006); 46 = Thompson and Townsend (multiple sources). [b]Data type: pii = prevalence and intensity of infection; I. Caught = individuals caught; F. of visits = frequency of visits; F. removed = fruits removed. [c]Number of organisms (species) with first function (#OA = number of anemone, plant, or predator species) or second function (#OB = number of fish, parasite, herbivore, pollinator, prey, or seed disperser species).


dx.doi.org/10.1021/ci400280n | J. Chem. Inf. Model. 2014, 54, 16−29

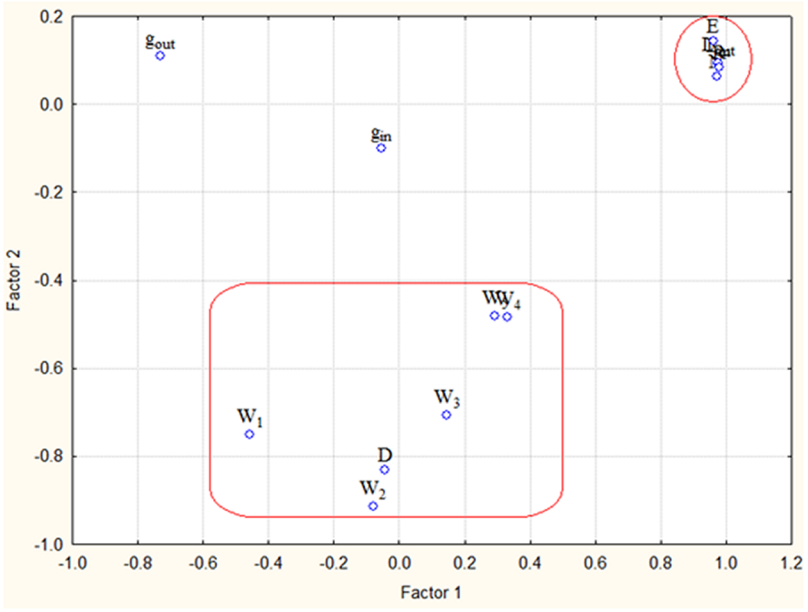



**Figure 2.** PCA of $W_k(j)$ values vs some classic parameters of MRNs of 43 organisms.

$$S_j = \sum_{k=0}^{5} a_k \cdot W_k(j) + \sum_{k=0}^{5} {}^1b_k \cdot [W_k(j) - W_k(j)_{Year.avg}] + \sum_{k=0}^{5} {}^2b_k \cdot [W_k(j) - W_k(j)_{Law.avg}] + c_0$$
$$= \sum_{k=0}^{5} a_k \cdot W_k(j) + \sum_{k=0}^{5} {}^1b_{gk} \cdot \Delta W_k(j)_{Year} + \sum_{k=0}^{5} {}^2b_{gk} \cdot \Delta W_k(j)_{Law} + c_0 \quad (8)$$

where $W_k(j)$, $W_k(j)_{Year.avg}$, and $W_k(j)_{Law.avg}$ are the Wiener-Markov node descriptor parameters of a given $j$-th Law and the average of these parameters for the given year (Year.avg) or for the same type of Financial Law (Law.avg). These parameters quantify information about the Legal regulations (Laws) of a given type introduced in the Spanish legal system at a given year with respect to the previous or successive $k^{th}$ laws approved.

In all cases, we used different statistical parameters to evaluate the statistical significance and validate the goodness-of-fit of ANN models: $n$ = number of cases, Specificity, and Sensitivity of both training and external validation series.[137]

## 3. RESULTS AND DISCUSSION

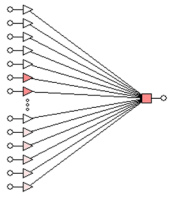
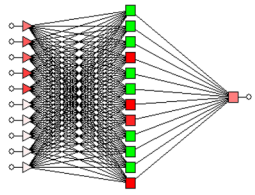
**3.1. MIANN-Wiener Models of MRNs.** The study of metabolic networks is very important in biology because many applications are directly built on the use of cellular metabolism.[138,139] Biotechnologists modify the cells and use them as cellular factories to produce antibiotics, industrial enzymes, antibodies, etc. In biomedicine, it is possible to cure metabolic diseases through a better understanding of the metabolic mechanisms and to control infections by making use of the metabolic differences between human beings and pathogens.[140] For example, the network topology-based approach has been used to uncover shared mechanisms in the study of disease comorbidity.[141] We carried out a Principal Component Analysis (PCA) of this data set (see Figure 2). We were able to explain 80% of all variance with only two principal components (pc). The first pc1, with an eigenvalue = 6.25, explains 48.1% of the variance, and the second component pc2, with an eigenvalue = 4.11, explains 31.6% of the variance. A third component pc3 was able to explain only 8.7% of the variance; consequently we discarded it. The results of this PCA are important to show that the new $W_k(j)$ indices codify useful structural information that is not trivially correlated with the information codified by other parameters. The PCA demonstrates that, in the case MRNs, the new Markov-Wiener indices codify different information compared with the classic ones.

Jeong et al.[132] showed that, despite significant variation in their individual constituents and pathways, metabolic networks have the same topological scaling properties and show striking similarities to the inherent organization of complex nonbiological systems. In any case, many pathways are not totally confirmed experimentally but have been computationally deduced using protein or gene alignment techniques. The idea follows more or less the following scheme: similar proteome → similar enzymes → similar metabolome. On the other hand, the experimental determination of the full metabolome including each metabolite and metabolite biotransformation pathways is not always an easy task. All these aspects determine the necessity of alignment-free techniques to assess network connectivity quality in existing models of metabolic pathway networks. Here we developed different MIANN models based on $W_k$ values to predict correct connectivity patterns of nodes in MRNs of 43 organisms belonging to different domains of the tree of life.

As seen in Table 3, the best MIANN model found presents very good values of Accuracy, Sensitivity, and Specificity for the recognition of links both in training and external validation series (see details in the Supporting Information SM1). The models were obtained using as input 15 descriptors: 5 Markov-Wiener node descriptors $W_k(j)$, 5 averages $W_k(j)_{g.avg}$, and 5 deviations $-\Delta W_k(j)_g$. The results obtained using the computer program STATISTICA show that the Multilayer Perceptron (MLP)[142] method fails to generate good prediction models, since it presents values of Specificity and Sensivity close to 50%.

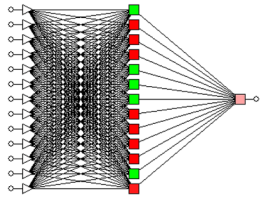
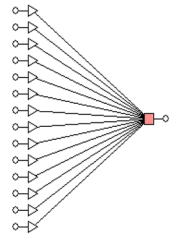

dx.doi.org/10.1021/ci400280n | J. Chem. Inf. Model. 2014, 54, 16−29

Table 3. MIANN Models of Metabolic Reaction Networks (MRNs)[a]

| ANN | Li | Li = 1 | Li = 0 | % | Pr. | % | Li = 1 | Li = 0 |
|---|---|---|---|---|---|---|---|---|
| LNN 15:15-1:1  | Li = 1 | 7276 | 1985 | 78.1 | Sn | 77.9 | 21917 | 6156 |
| | Li = 0 | 2044 | 7066 | 78.1 | Sp | 77.6 | 6227 | 21329 |
| MLP 2:2-11-1:1  | Li = 1 | 4669 | 4559 | 50.1 | Sn | 49.7 | 13990 | 13856 |
| | Li = 0 | 4651 | 4492 | 49.6 | Sp | 49.6 | 14154 | 13629 |

[a]Pr. = Parameter, Sp = Specificity, Sn = Sensitivity. Columns: Observed classifications; Rows: Predicted classifications; MLP = Multilayer Perceptron; LNN = Linear Neural Network.

Table 4. MIANN Models of the IWDB Complex Networks[a]

| ANN | Li | Li = 1 | Li = 0 | % | Pr. | % | Li = 1 | Li = 0 |
|---|---|---|---|---|---|---|---|---|
| MLP 13:13-13-1:1  | Li = 1 | 4570 | 547 | 91.1 | Sn | 90.5 | 1363 | 194 |
| | Li = 0 | 449 | 4346 | 88.8 | Sp | 88.1 | 143 | 1437 |
| LNN 14:14-1:1  | Li = 1 | 3326 | 1710 | 66.3 | Sn | 66.1 | 995 | 603 |
| | Li = 0 | 1693 | 3183 | 65.1 | Sp | 63.0 | 511 | 1028 |

[a]Pr. = Parameter, Sp = Specificity, Sn = Sensitivity. Columns: Observed classifications; Rows: Predicted classifications; MLP = Multilayer Perceptron; LNN = Linear Neural Network.

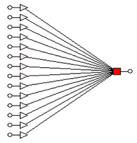
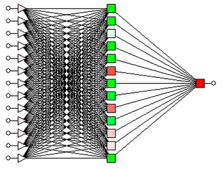
On the other hand, the LNN based on 15 descriptors (LNN 15:15−1:1) is able to classify correctly 78.1% of the cases, with a sensitivity of 77.9% and a specificity of 77.6%. The LNN is equivalent to a LDA equation, the simplest type of classification model.

**3.2. MIANN-Wiener Models of IWDBNs.** We tested different MIANN models with linear (LNN) and nonlinear (ANN) forms. The results are presented in Table 4, and the details can be found in the Supporting Information SM2. In the case of the IWDBNs, the best classification model is obtained with the MLP classifier based on 13 input descriptors and 13 neurons in the hidden layer (MLP 13:13−13−1:1). This model can classify 91.1% of the nodes with a sensitivity of 90.5% and specificity of 88.8%. Unlike the case of the MRNs, the LNN is not able to classify the IWDBN's nodes with accuracy (<67%). Thus, it can be observed that, compared with the MRNs, the IWDBNs contain more complex information for the classification of the connectivity between nodes. The IWDBNs need complex classifiers such as MLPs in comparison with the MRNs that can be processed using the simpler LNNs.

**3.3. MIANN-Wiener Models of SFLN.** The use of network analysis methods in social sciences began in 1930 and today are widely used.[143] However, the application of these methods in legal studies is still at the beginning.[144−146] The network tools can illustrate the interrelation between different laws and help to understand their consequences on the society. We have used the list of the financial laws to construct the studied network. The best models found are presented in Table 5. We tested different MIANN models with linear (LNN) and nonlinear (ANN) forms. These MIANN models behave like time series embedded within a complex network. This is due to the fact it predicts the recurrence of the Spanish law system to a financial regulation of class c when the social and economical conditions change at time $t_{i+1}$ when a known class of regulation has been


dx.doi.org/10.1021/ci400280n | J. Chem. Inf. Model. 2014, 54, 16−29

Table 5. MIANN Time Series Model of Spanish Financial Law Network (SFLN)[a]

| ANN | Li | Li = 1 | Li = 0 | % | Pr. | % | Li = 1 | Li = 0 |
|---|---|---|---|---|---|---|---|---|
| LNN 14:14-1:1  | Li = 1 | 125 | 41 | 86.2 | Sn | 87.4 | 370 | 156 |
| | Li = 0 | 18 | 298 | 85.4 | Sp | 87.9 | 59 | 914 |
| MLP 14:14-14:1  | Li = 1 | 119 | 54 | 85.3 | Sn | 83.2 | 366 | 129 |
| | Li = 0 | 24 | 285 | 87.9 | Sp | 84.1 | 63 | 941 |

[a]Pr. = Parameter, Sp = Specificity, Sn = Sensitivity. Columns: Observed classifications; Rows: Predicted classifications; MLP = Multilayer Perceptron; LNN = Linear Neural Network.

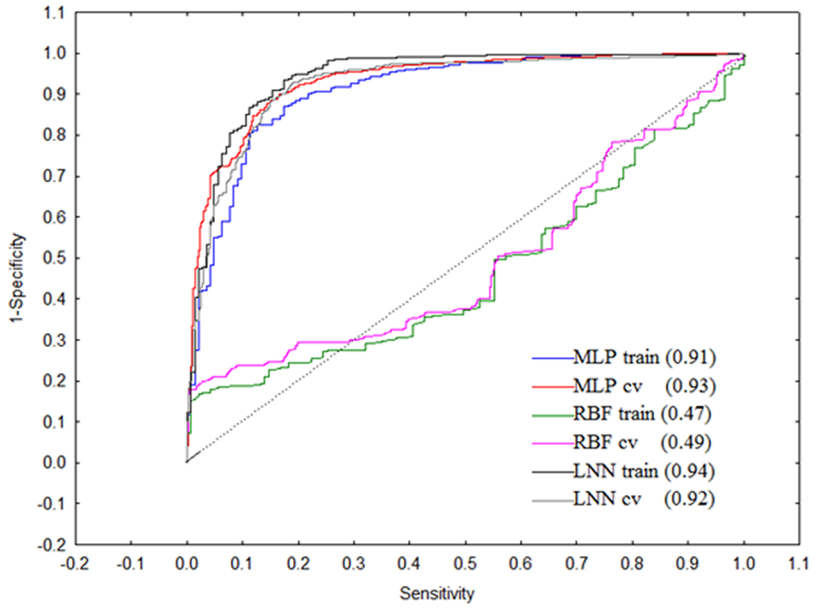



**Figure 3.** ROC curve analysis of the SFLN.

used in the past at time $t_i$. The best model correctly reconstructed the network of the historical record for the Spanish financial system with high Accuracy, Specificity, and Sensitivity (see Table 5). Detailed results for each case (including codes, classification, probability, and node descriptors values) are given in the Supporting Information SM3. In the case of the SFLN, there is no clear difference of Accuracy, Specificity, and Sensitivity between the two models studied (LNN and MLP). In this situation we can apply the Occam's razor and choose the LNN model, which is the simplest. However, to be more certain of this choice, we decided to carry out a ROC curve analysis. The AUROC values (Area Under Receiver Operating Characteristic) and the ROC curves for three different MIANN models (MLP, LNN, and RBF-or Radial Basis Function-) are presented in Figure 3. We show separately the values for training and validation series. The values obtained confirm that the LNN model based on 14 descriptors is the best found in this case, with a correct classification of 86.2%, sensitivity of 87.4%, and specificity of 87.9%. The best RBF classifier, which is based on only one descriptor, is not able to classify the SFLN's nodes.

## 4. CONCLUSIONS

This work introduces a new type of node descriptors, the Markov-Wiener node descriptors of order $k^{th}$ ($W_k$), higher-order analogues of the classic Wiener index, a graph invariant widely used in chemoinformatics. The new node descriptors are used to search for classification models able to discriminate the correct node connectivity patterns from the incorrect random patterns. The classifiers are obtained by using Artificial Intelligence algorithms called Artificial Neural Networks (ANNs). This mixture of Markov node descriptors and ANNs is presented as the MIANN method. The classifiers, based on LNN and MLP, showed good values of Sensitivity/Specificity (%) for the studied networks: MRNs (78/78), IWDBNs (90/88), and SFLN (86/84).

The use of the new Markov-Wiener node descriptors demonstrates that it is possible to carry out a theoretical re-evaluation of the connectivity in known complex networks (collation) as a fast alternative to the high-cost experimental re-evaluation of all the links of the studied network.


dx.doi.org/10.1021/ci400280n | J. Chem. Inf. Model. 2014, 54, 16−29

## ■ ASSOCIATED CONTENT

### Supporting Information

SM1, SM2, and SM3. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION


### Corresponding Author

*Phone: +34 981 167 000, Ext. 1302. Fax: +34 981 167 160. E-mail: crm.publish@gmail.com.


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

C. R. Munteanu acknowledges the Isidro Parga Pondal Program, funded by Xunta de Galicia, Spain and the European Social Fund (ESF), for partial financial support.

## ■ REFERENCES


(1) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: 2002.

(2) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

(3) Wiener, H. Correlation of heats of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons. *J. Am. Chem. Soc.* **1947**, *69*, 2636−2638.

(4) Wiener, H. Relation of the physical properties of the isomeric alkanes to molecular structure. Surface tension, specific dispersion, and critical solution temperature in aniline. *J. Phys. Colloid Chem.* **1948**, *52*, 1082−1089.

(5) Wiener, H. Vapor pressure−temperature relationships among the branched paraffin hydrocarbons. *J. Phys. Colloid Chem.* **1948**, *52*, 425−430.

(6) Hosoya, H. Topological index, a newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339.

(7) Hosoya, H. Mathematical meaning and importance of the topological index Z. *Croat. Chem. Acta* **2007**, *80*, 239−249.

(8) Consonni, V.; Todeschini, R. Molecular descriptors. In *Recent advances in QSAR studies: Methods and applications*; Puzyn, T., Leszczynski, J., Cronin, M. T. D., Eds.; Springer: 2010; Chapter 3, pp 29−102.

(9) Harary, F. Status and contrastatus. *Sociometry* **1959**, *22*, 23−43.

(10) Diudea, M. V.; Gutman, I. Wiener-type topological indices. *Croat. Chem. Acta* **1998**, *71*, 21−51.

(11) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH: Weinheim, 2009; p 1257.

(12) Balaban, A. T. From chemical graphs to 3D molecular modeling. In *From chemical topology to three-dimensional geometry*; Balaban, A. T., Ed.; Plenum Publishers: New York, 1997, p 420.

(13) Mandloi, M.; Sikarwar, A.; Sapre, N. S.; Karmarkar, S.; Khadikar, P. V. A comparative QSAR study using Wiener, Szeged, and molecular connectivity indices. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 57−62.

(14) Lukovits, I. Decomposition of the Wiener topological index. Application to drug-receptor interactions. *J. Chem. Soc., Perkin Trans. 2* **1988**, 1667−1671.

(15) Lukovits, I. Correlation between components of the Wiener index and partition coefficients of hydrocarbons. *Int. J. Quantum Chem.* **1992**, *44*, 217−223.

(16) Mendiratta, S.; Madan, A. K. Structure-activity study on antiviral 5-vinylpyrimidine nucleoside analogs using Wiener's topological index. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 867−871.

(17) Agrawal, V. K.; Srivastava, R.; Khadikar, P. V. QSAR studies on some antimalarial sulfonamides. *Bioorg. Med. Chem.* **2001**, *9*, 3287−3293.

(18) Sardana, S.; Madan, A. K. Application of graph theory: Relationship of molecular connectivity index, Wiener's index and eccentric connectivity index with diuretic activity. *MATCH* **2001**, *43*, 85−98.

(19) Sardana, S.; Madan, A. K. Predicting anticonvulsant activity of benzamides/benzylamines: computational approach using topological descriptors. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 545−550.

(20) Gupta, S.; Singh, M.; Madan, A. K. Application of graph theory: Relationship of eccentric connectivity Index and Wiener's Index with anti-inflammatory activity. *J. Math. Anal. Appl.* **2002**, *266*, 259−268.

(21) Bajaj, S.; Sambi, S. S.; Madan, A. K. Predicting anti-HIV activity of phenethylthiazolethiourea (PETT) analogs: computational approach using Wiener's topochemical index. *J. Mol. Struct.* **2004**, *684*, 197−203.

(22) Bajaj, S.; Sambi, S. S.; Madan, A. K. Topological models for prediction of anti-HIV activity of acylthiocarbamates. *Bioorg. Med. Chem.* **2005**, *13*, 3263−3268.

(23) Bajaj, S.; Sambi, S. S.; Madan, A. K. Topological models for prediction of anti-inflammatory activity of N-arylanthranilic acids. *Bioorg. Med. Chem. Lett.* **2004**, *12*, 3695−3701.

(24) Bajaj, S.; Sambi, S. S.; Madan, A. K. Topochemical models for prediction of anti-tumor activity of 3-aminopyrazoles. *Chem. Pharm. Bull. (Tokyo)* **2005**, *53*, 611−615.

(25) Kumar, V.; Madan, A. K. Application of graph theory: prediction of glycogen synthase kinase-3 beta inhibitory activity of thiadiazolidinones as potential drugs for the treatment of Alzheimer's disease. *Eur. J. Pharm. Sci.* **2005**, *24*, 213−218.

(26) Lather, V.; Madan, A. K. Topological model for the prediction of MRP1 inhibitory activity of pyrrolopyrimidines and templates derived from pyrrolopyrimidine. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 4967−4972.

(27) Bornholdt, S.; Schuster, H. G. *Handbook of Graphs and Complex Networks: From the Genome to the Internet*; WILEY-VCH: Wheinheim, 2003.

(28) Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; Hwang, D. U. Complex networks: Structure and dynamics. *Phys. Rep.* **2006**, *424*, 175−308.

(29) Dehmer, M.; Emmert-Streib, F. *Analysis of complex networks: from biology to linguistics*; Wiley-Blackwell: Weinheim, 2009; p 462.

(30) Newman, M. The structure and function of complex networks. *SIAM Rev.* **2003**, *45*, 167−256.

(31) Thomas, S.; Bonchev, D. A survey of current software for network analysis in molecular biology. *Hum. Genomics* **2010**, *4*, 353−360.

(32) Bonchev, D.; Buck, G. A. From molecular to biological structure and back. *J. Chem Inf. Model.* **2007**, *47*, 909−917.

(33) Bonchev, D.; Rouvray, D. H. *Complexity in Chemistry, Biology, and Ecology*; Springer Science+Business Media, Inc.: New York, 2005.

(34) Bonchev, D. Complexity analysis of yeast proteome network. *Chem. Biodivers.* **2004**, *1*, 312−326.

(35) Bonchev, D. On the complexity of directed biological networks. *SAR QSAR Environ. Res.* **2003**, *14*, 199−214.

(36) Gonzalez-Diaz, H. QSAR and complex networks in pharmaceutical design, microbiology, parasitology, toxicology, cancer, and neurosciences. *Curr. Pharm. Des.* **2010**, *16*, 2598−2600.

(37) Gonzalez-Diaz, H. Network topological indices, drug metabolism, and distribution. *Curr. Drug. Metab* **2010**, *11*, 283−284.

(38) 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. Pharmaceutics* **2009**, *6*, 825−835.

(39) Duardo-Sanchez, A.; Patlewicz, G.; González-Díaz, H. A review of network topological indices from chem-bioinformatics to legal sciences and back. *Curr. Bioinf.* **2011**, *6*, 53−70.

(40) Puzyn, T.; Leszczynski, J.; Cronin, M. T. D. *Recent Advances in QSAR Studies: Methods and applications*; Springer: London, 2010; p 423.

(41) González-Díaz, H.; Munteanu, C. R. *Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks*; Transworld Research Network: Kerala, India, 2010.

(42) González-Díaz, H.; González-Díaz, Y.; Santana, L.; Ubeira, F. M.; Uriarte, E. Proteomics, networks and connectivity indices. *Proteomics* **2008**, *8*, 750−778.
(43) Gonzalez-Diaz, H.; Duardo-Sanchez, A.; Ubeira, F. M.; Prado-Prado, F.; Perez-Montoto, L. G.; Concu, R.; Podda, G.; Shen, B. Review of MARCH-INSIDE & complex networks prediction of drugs: ADMET, anti-parasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers. *Curr. Drug. Metab.* **2010**, *11*, 379−406.
(44) Gonzalez-Diaz, H.; Prado-Prado, F.; Garcia-Mera, X.; Alonso, N.; Abeijon, P.; Caamano, O.; Yanez, M.; Munteanu, C. R.; Pazos, A.; Dea-Ayuela, M. A.; Gomez-Munoz, M. T.; Garijo, M. M.; Sansano, J.; Ubeira, F. M. MIND-BEST: Web server for drugs and target discovery; design, synthesis, and assay of MAO-B inhibitors and theoretical-experimental study of G3PDH protein from Trichomonas gallinae. *J. Proteome Res.* **2011**, *10*, 1698−1718.
(45) Rodriguez-Soca, Y.; Munteanu, C. R.; Dorado, J.; Pazos, A.; Prado-Prado, F. J.; Gonzalez-Diaz, H. Trypano-PPI: a web server for prediction of unique targets in trypanosome proteome by using electrostatic parameters of protein-protein interactions. *J. Proteome Res.* **2010**, *9*, 1182−1190.
(46) Gonzalez-Diaz, H.; Romaris, F.; Duardo-Sanchez, A.; Perez-Montoto, L. G.; Prado-Prado, F.; Patlewicz, G.; Ubeira, F. M. Predicting drugs and proteins in parasite infections with topological indices of complex networks: theoretical backgrounds, applications, and legal issues. *Curr. Pharm. Des.* **2010**, *16*, 2737−2764.
(47) Gonzalez-Diaz, H.; Prado-Prado, F. J.; Garcia-Mera, X.; Alonso, N.; Abeijon, P.; Caamano, O.; Yanez, M.; Munteanu, C. R.; Pazos Sierra, A.; Dea-Ayuela, M. A.; Gomez-Munoz, M. T.; Garijo, M. M.; Sansano, J.; Ubeira, F. M. MIND-BEST: web server for drugs & target discovery; design, synthesis, and assay of MAO-B inhibitors and theoretic-experimental study of G3PD protein from Trichomona gallineae. *J. Proteome Res.* **2010**, *10*, 1698−1718.
(48) Munteanu, C. R.; Vazquez, J. M.; Dorado, J.; Sierra, A. P.; Sanchez-Gonzalez, A.; Prado-Prado, F. J.; Gonzalez-Diaz, H. Complex network spectral moments for ATCUN motif DNA cleavage: first predictive study on proteins of human pathogen parasites. *J. Proteome Res.* **2009**, *8*, 5219−5228.
(49) 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*, 4372−4382.
(50) Aguero-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*, 2122−2128.
(51) 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*, 6740−6751.
(52) Gonzalez-Diaz, H.; Prado-Prado, F.; Ubeira, F. M. Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. *Curr. Top. Med. Chem.* **2008**, *8*, 1676−1690.
(53) Aguero-Chapin, G.; Gonzalez-Diaz, H.; de la Riva, G.; Rodriguez, E.; Sanchez-Rodriguez, A.; Podda, G.; Vazquez-Padron, R. I. MMM-QSAR recognition of ribonucleases without alignment: comparison with an HMM model and isolation from Schizosaccharomyces pombe, prediction, and experimental assay of a new sequence. *J. Chem. Inf. Model.* **2008**, *48*, 434−448.
(54) Aguero-Chapin, G.; Antunes, A.; Ubeira, F. M.; Chou, K. C.; Gonzalez-Diaz, H. Comparative study of topological indices of macro/supramolecular RNA complex networks. *J. Chem. Inf. Model.* **2008**, *48*, 2265−2277.
(55) 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*, 1025−1039.
(56) 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*, 715−723.
(57) González-Díaz, H.; Marrero, Y.; Hernandez, I.; Bastida, I.; Tenorio, E.; Nasco, O.; Uriarte, E.; Castanedo, N.; Cabrera, M. A.; Aguila, E.; Marrero, O.; Morales, A.; Perez, M. 3D-MEDNEs: an alternative "in silico" technique for chemical research in toxicology. 1. prediction of chemically induced agranulocytosis. *Chem. Res. Toxicol.* **2003**, *16*, 1318−1327.
(58) Berca, M. N.; Duardo-Sanchez, A.; González-Díaz, H.; Pazos, A.; Munteanu, C. R. Markov entropy for biology, parasitology, linguistic, technology, social and law networks. In *Complex Network Entropy: From Molecules to Biology, Parasitology, Technology, Social, Legal, and Neurosciences*; González-Díaz, H., Prado-Prado, F. J., García-Mera, X., Eds.; Transworld Research Network: Kerala, India, 2011; Chapter 10, pp 127−142.
(59) Aguiar-Pulido, V.; Seoane-Fernańdez, J. A.; Freire-Veiga, A. M.; Gonzalez-Diaz, H.; Duardo-Sanchez, A.; Dorado, J.; Pazos, A.; Munteanu, C. R. New Markov-Randic Centralities for Computational methods of Biology, Parasitology, Technology, Social and Law networks. Proceedings of ICCMSE 2010, Kos, Greece. AIP Conference Proceedings, Melville, NY, USA, accepted (2014).
(60) González-Díaz, H.; Riera-Fernández, P.; Pazos, A.; Munteanu, C. R. The Rücker-Markov invariants of complex bio-systems: applications in parasitology and neuroinformatics. *Biosystems* **2013**, *111*, 199−207.
(61) Riera-Fernandez, P.; Munteanu, C. R.; Dorado, J.; Martin-Romalde, R.; Duardo-Sanchez, A.; Gonzalez-Diaz, H. From chemical graphs in computer-aided drug design to general Markov-Galvez indices of drug-target, proteome, drug-parasitic disease, technological, and social-legal networks. *Curr. Comput.-Aided Drug Des.* **2011**, *7*, 315−337.
(62) Gonzalez-Diaz, H.; Riera-Fernandez, P. New Markov-autocorrelation indices for re-evaluation of links in chemical and biological complex networks used in metabolomics, parasitology, neurosciences, and epidemiology. *J. Chem. Inf. Model.* **2012**, *52*, 3331−3340.
(63) Riera-Fernández, P.; Munteanu, C. R.; 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. *Curr. Bioinf.* **2011**, *6*, 94−121.
(64) Estrada, E.; Delgado, E. J.; Alderete, J. B.; Jaña, G. A. Quantum-connectivity descriptors in modeling solubility of environmentally important organic compounds. *J. Comput. Chem.* **2004**, *25*, 1787−1796.
(65) Besalu, E.; Girones, X.; Amat, L.; Carbo-Dorca, R. Molecular quantum similarity and the fundamentals of QSAR. *Acc. Chem. Res.* **2002**, *35*, 289−295.
(66) Rincon, D. A.; Cordeiro, M. N.; Mosquera, R. A. On the electronic structure of cocaine and its metabolites. *J. Phys. Chem. A* **2009**, *113*, 13937−13942.
(67) Mandado, M.; Gonzalez-Moa, M. J.; Mosquera, R. A. Chemical graph theory and n-center electron delocalization indices: a study on polycyclic aromatic hydrocarbons. *J. Comput. Chem.* **2007**, *28*, 1625−1633.
(68) Gonzalez-Diaz, H.; Gonzalez-Diaz, Y.; Santana, L.; Ubeira, F. M.; Uriarte, E. Proteomics, networks and connectivity indices. *Proteomics* **2008**, *8*, 750−778.
(69) Helguera, A. M.; Combes, R. D.; Gonzalez, M. P.; Cordeiro, M. N. Applications of 2D descriptors in drug design: a DRAGON tale. *Curr. Top. Med. Chem.* **2008**, *8*, 1628−1655.
(70) Casanola-Martin, G. M.; Marrero-Ponce, Y.; Khan, M. T.; Ather, A.; Khan, K. M.; Torrens, F.; Rotondo, R. Dragon method for finding novel tyrosinase inhibitors: Biosilico identification and experimental in vitro assays. *Eur. J. Med. Chem.* **2007**, *42*, 1370−1381.
(71) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov,

N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual computational chemistry laboratory–design and description. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 453−463.
(72) Marzaro, G.; Chilin, A.; Guiotto, A.; Uriarte, E.; Brun, P.; Castagliuolo, I.; Tonus, F.; Gonzalez-Diaz, H. Using the TOPS-MODE approach to fit multi-target QSAR models for tyrosine kinases inhibitors. *Eur. J. Med. Chem.* **2011**, *46*, 2185−2192.
(73) Vilar, S.; Estrada, E.; Uriarte, E.; Santana, L.; Gutierrez, Y. In silico studies toward the discovery of new anti-HIV nucleoside compounds through the use of TOPS-MODE and 2D/3D connectivity indices. 2. Purine derivatives. *J. Chem. Inf. Model.* **2005**, *45*, 502−514.
(74) Estrada, E.; Quincoces, J. A.; Patlewicz, G. Creating molecular diversity from antioxidants in Brazilian propolis. Combination of TOPS-MODE QSAR and virtual structure generation. *Mol. Divers.* **2004**, *8*, 21−33.
(75) Estrada, E.; Gonzalez, H. What are the limits of applicability for graph theoretic descriptors in QSPR/QSAR? Modeling dipole moments of aromatic compounds with TOPS-MODE descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 75−84.
(76) Marrero-Ponce, Y.; Castillo-Garit, J. A.; Olazabal, E.; Serrano, H. S.; Morales, A.; Castañedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Jorge, E.; del Valle, A.; Torrens, F.; Castro, E. A. TOMOCOMD-CARDD, a novel approach for computer-aided 'rational' drug design: I. Theoretical and experimental assessment of a promising method for computational screening and in silico design of new anthelmintic compounds. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 615−634.
(77) Marrero-Ponce, Y.; Medina-Marrero, R.; Castro, A. E.; Ramos de Armas, R.; González-Díaz, H.; Romero-Zaldivar, V.; Torrens, F. Protein quadratic indices of the "macromolecular pseudograph's $\alpha$-carbon atom adjacency matrix". 1. Prediction of Arc repressor alanine-mutant's stability. *Molecules* **2004**, *9*, 1124−1147.
(78) Katritzky, A. R.; 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*, 3453−3457.
(79) Katritzky, A. R.; Perumal, S.; Petrukhin, R.; Kleinpeter, E. Codessa-based theoretical QSPR model for hydantoin HPLC-RT lipophilicities. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 569−574.
(80) Vilar, S.; Cozza, G.; Moro, S. Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. *Curr. Top. Med. Chem.* **2008**, *8*, 1555−1572.
(81) Junker, B. H.; Koschutzki, D.; Schreiber, F. Exploration of biological network centralities with CentiBiN. *BMC Bioinf.* **2006**, *7*, 219.
(82) Batagelj, V.; Mrvar, A. Pajek— analysis and visualization of large networks. *Lect. Notes Comput. Sci.* **2002**, *2265*, 477−478.
(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) Frank, E.; Hall, M.; Trigg, L.; Holmes, G.; Witten, I. H. Data mining in bioinformatics using Weka. *Bioinformatics* **2004**, *20*, 2479−2481.
(85) Bisson, W. H. Editorial: Computational chemogenomics in drug design and discovery. *Curr. Top. Med. Chem.* **2012**, *12*, 1867−1868.
(86) Cordeiro, M. N.; Speck-Planche, A. Editorial: Computer-aided drug design, synthesis and evaluation of new anti-cancer drugs. *Curr. Top. Med. Chem.* **2012**, *12*, 2703−2704.
(87) Prado-Prado, F.; Garcia-Mera, X. Editorial: QSAR models for computer-aided drug design and molecular docking for disorders of the central nervous system and other diseases. *Curr. Top. Med. Chem.* **2012**, *12*, 1731−1733.
(88) Gonzalez-Diaz, H. *Quantitative studies on Structure-Activity and Structure-Property Relationships (QSAR/QSPR)*; 2008; Vol. *8*, p 1554.
(89) Gonzalez-Diaz, H. Editorial: QSAR/QSPR models as enabling technologies for drug & targets discovery in: medicinal chemistry, microbiology-parasitology, neurosciences, bioinformatics, proteomics and other biomedical sciences. *Curr. Top. Med. Chem.* **2012**, *12*, 799−801.
(90) Alderson, R. G.; De Ferrari, L.; Mavridis, L.; McDonagh, J. L.; Mitchell, J. B.; Nath, N. Enzyme informatics. *Curr. Top. Med. Chem.* **2012**, *12*, 1911−1923.
(91) Bisson, W. H. Drug repurposing in chemical genomics: can we learn from the past to improve the future? *Curr. Top. Med. Chem.* **2012**, *12*, 1883−1888.
(92) Castillo-Garit, J. A.; Abad, C.; Rodriguez-Borges, J. E.; Marrero-Ponce, Y.; Torrens, F. A review of QSAR studies to discover new drug-like compounds actives against leishmaniasis and trypanosomiasis. *Curr. Top. Med. Chem.* **2012**, *12*, 852−865.
(93) Cedeno, W.; Alex, S.; Jaeger, E. P.; Agrafiotis, D. K.; Lobanov, V. S. An integrated data management framework for drug discovery–from data capturing to decision support. *Curr. Top. Med. Chem.* **2012**, *12*, 1237−1242.
(94) Chatterjee, A. K.; Yeung, B. K. Back to the future: lessons learned in modern target-based and whole-cell lead optimization of antimalarials. *Curr. Top. Med. Chem.* **2012**, *12*, 473−483.
(95) Chen, J.; Wang, Y.; Guo, D.; Shen, B. A systems biology perspective on rational design of peptide vaccine against virus infections. *Curr. Top. Med. Chem.* **2012**, *12*, 1310−1319.
(96) Cordero, F.; Beccuti, M.; Donatelli, S.; Calogero, R. A. Large disclosing the nature of computational tools for the analysis of next generation sequencing data. *Curr. Top. Med. Chem.* **2012**, *12*, 1320−1330.
(97) Dave, K.; Lahiry, A. Conotoxins: review and docking studies to determine potentials of conotoxin as an anticancer drug molecule. *Curr. Top. Med. Chem.* **2012**, *12*, 845−851.
(98) Dave, K.; Panchal, H. Review on chemogenomics approach: interpreting antagonist activity of secreted frizzled-related protein 1 in glaucoma disease with in-silico docking. *Curr. Top. Med. Chem.* **2012**, *12*, 1834−1842.
(99) Faivre, C.; Barbolosi, D.; Iliadis, A. A new model for determining the MTD during phase-I trials in pediatric oncology. *Curr. Top. Med. Chem.* **2012**, *12*, 1660−1664.
(100) Garcia, I.; Fall, Y.; Gomez, G. Review of synthesis, biological assay, and QSAR studies of HMGR inhibitors. *Curr. Top. Med. Chem.* **2012**, *12*, 895−919.
(101) Jayadeepa, R. M.; Niveditha, M. S. Computational approaches to screen candidate ligands with anti- Parkinson's activity using R programming. *Curr. Top. Med. Chem.* **2012**, *12*, 1807−1814.
(102) Khan, M. T.; Mischiati, C.; Ather, A.; Ohyama, T.; Dedachi, K.; Borgatti, M.; Kurita, N.; Gambari, R. Structure-based analysis of the molecular recognitions between HIV-1 TAR-RNA and transcription factor nuclear factor-kappaB (NFkB). *Curr. Top. Med. Chem.* **2012**, *12*, 814−827.
(103) Kobe, B.; Boden, M. Computational modelling of linear motif-mediated protein interactions. *Curr. Top. Med. Chem.* **2012**, *12*, 1553−1561.
(104) Kramer, C.; Lewis, R. QSARs, data and error in the modern age of drug discovery. *Curr. Top. Med. Chem.* **2012**, *12*, 1896−1902.
(105) Kufareva, I.; Chen, Y. C.; Ilatovskiy, A. V.; Abagyan, R. Compound activity prediction using models of binding pockets or ligand properties in 3D. *Curr. Top. Med. Chem.* **2012**, *12*, 1869−1882.
(106) Lin, J. H. Target prediction of small molecules with information of key molecular interactions. *Curr. Top. Med. Chem.* **2012**, *12*, 1903−1910.
(107) Luan, F.; Borges, F.; Cordeiro, M. N. Recent advances on A(3) adenosine receptor antagonists by QSAR tools. *Curr. Top. Med. Chem.* **2012**, *12*, 878−894.
(108) Mortier, J.; Rakers, C.; Frederick, R.; Wolber, G. Computational tools for in silico fragment-based drug design. *Curr. Top. Med. Chem.* **2012**, *12*, 1935−1943.
(109) Ortore, G.; Tuccinardi, T.; Martinelli, A. Computational studies on translocator protein (TSPO) and its ligands. *Curr. Top. Med. Chem.* **2012**, *12*, 352−359.
(110) Popelier, P. New insights in atom-atom interactions for future drug design. *Curr. Top. Med. Chem.* **2012**, *12*, 1924−1934.

(111) Prado-Prado, F.; Garcia-Mera, X.; Escobar, M.; Alonso, N.; Caamano, O.; Yanez, M.; Gonzalez-Diaz, H. 3D MI-DRAGON: new model for the reconstruction of US FDA drug- target network and theoretical-experimental studies of inhibitors of rasagiline derivatives for AChE. *Curr. Top. Med. Chem.* **2012**, *12*, 1843−1865.
(112) Riera-Fernandez, I.; Martin-Romalde, R.; Prado-Prado, F. J.; Escobar, M.; Munteanu, C. R.; Concu, R.; Duardo-Sanchez, A.; Gonzalez-Diaz, H. From QSAR models of drugs to complex networks: state-of-art review and introduction of new Markov-spectral moments Indices. *Curr. Top. Med. Chem.* **2012**, *12*, 927−960.
(113) Saladino, G.; Gervasio, F. L. New insights in protein kinase conformational dynamics. *Curr. Top. Med. Chem.* **2012**, *12*, 1889−1895.
(114) Sharma, N.; Ethiraj, K. R.; Yadav, M.; Nayarisseri, S. A.; Chaurasiya, M.; Vankudavath, R. N.; Rao, K. R. Identification of LOGP values and electronegativities as structural insights to model inhibitory activity of HIV-1 capsid inhibitors - A SVM and MLR aided QSAR studies. *Curr. Top. Med. Chem.* **2012**, *12*, 1763−1774.
(115) Speck-Planche, A.; Kleandrova, V. V. QSAR and molecular docking techniques for the discovery of potent monoamine oxidase B inhibitors: computer-aided generation of new rasagiline bioisosteres. *Curr. Top. Med. Chem.* **2012**, *12*, 1734−1747.
(116) Van Calenbergh, S.; Pochet, S.; Munier-Lehmann, H. Drug design and identification of potent leads against mycobacterium tuberculosis thymidine monophosphate kinase. *Curr. Top. Med. Chem.* **2012**, *12*, 694−705.
(117) Zhang, T.; Zhao, M.; Pang, Y.; Zhang, W.; Angela Liu, L.; Wei, D. Q. Recent progress on bioinformatics, functional genomics, and metabolomics research of cytochrome P450 and its impact on drug discovery. *Curr. Top. Med. Chem.* **2012**, *12*, 1346−1355.
(118) Gonzalez-Diaz, H.; Romaris, F.; Duardo-Sanchez, A.; Perez-Montoto, L. G.; Prado-Prado, F.; Patlewicz, G.; Ubeira, F. M. Predicting drugs and proteins in parasite infections with topological indices of complex networks: theoretical backgrounds, applications, and legal issues. *Curr. Pharm. Des.* **2010**, *16*, 2737−2764.
(119) Caballero, J.; Fernandez, M. Artificial neural networks from MATLAB in medicinal chemistry. Bayesian-regularized genetic neural networks (BRGNN): application to the prediction of the antagonistic activity against human platelet thrombin receptor (PAR-1). *Curr. Top. Med. Chem.* **2008**, *8*, 1580−1605.
(120) Duardo-Sanchez, A.; Patlewicz, G.; Lopez-Diaz, A. Current topics on software use in medicinal chemistry: intellectual property, taxes, and regulatory issues. *Curr. Top. Med. Chem.* **2008**, *8*, 1666−1675.
(121) Ivanciuc, O. Weka machine learning for predicting the phospholipidosis inducing potential. *Curr. Top. Med. Chem.* **2008**, *8*, 1691−1709.
(122) Wang, J. F.; Wei, D. Q.; Chou, K. C. Drug candidates from traditional chinese medicines. *Curr. Top. Med. Chem.* **2008**, *8*, 1656−1665.
(123) Gonzalez-Diaz, 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*, 1015−1029.
(124) Bhattacharjee, B.; Jayadeepa, R. M.; Banerjee, S.; Joshi, J.; Middha, S. K.; Mole, J. P.; Samuel, J. Review of complex network and gene ontology in pharmacology approaches: Mapping natural compounds on potential drug target colon cancer network. *Curr. Bioinf.* **2011**, *6*, 44−52.
(125) Dave, K.; Banerjee, A. Bioinformatics analysis of functional relations between CNPs regions. *Curr. Bioinf.* **2011**, *6*, 122−128.
(126) García, I.; Fall, Y.; Gómez, G. Trends in bioinformatics and chemoinformatics of vitamin D analogues and their protein targets. *Curr. Bioinf.* **2011**, *6*, 16−24.
(127) Ivanciuc, T.; Ivanciuc, O.; Klein, D. J. Network-QSAR with reaction poset quantitative superstructure-activity relationships (QSSAR) for PCB chromatographic properties. *Curr. Bioinf.* **2011**, *6*, 25−34.
(128) Prado-Prado, F.; Escobar-Cubiella, M.; García-Mera, X. Review of bioinformatics and QSAR studies of $\beta$-secretase inhibitors. *Curr. Bioinf.* **2011**, *6*, 3−15.
(129) Wan, S. B.; Hu, L. L.; Niu, S.; Wang, K.; Cai, Y. D.; Lu, W. C.; Chou, K. C. Identification of multiple subcellular locations for proteins in budding yeast. *Curr. Bioinf.* **2011**, *6*, 71−80.
(130) Gonzalez-Diaz, H.; Duardo-Sanchez, A.; Ubeira, F. M.; Prado-Prado, F.; Perez-Montoto, L. G.; Concu, R.; Podda, G.; Shen, B. Review of MARCH-INSIDE & complex networks prediction of drugs: ADMET, anti-parasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers. *Curr. Drug Metab.* **2010**, *11*, 379−406.
(131) Gonzalez-Diaz, H.; Arrasate, S.; Sotomayor, N.; Lete, E.; Munteanu, C. R.; Pazos, A.; Besada-Porto, L.; Ruso, J. M. MIANN models in medicinal, physical and organic chemistry. *Curr. Top. Med. Chem.* **2013**, *13*, 619−641.
(132) Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z. N.; Barabasi, A. L. The large-scale organization of metabolic networks. *Nature* **2000**, *407*, 651−654.
(133) Riera-Fernandez, P.; Munteanu, C. R.; Pedreira-Souto, N.; Martin-Romalde, R.; Duardo-Sanchez, A.; Gonzalez-Diaz, H. Definition of Markov-Harary invariants and review of classic topological indices and databases in biology, parasitology, technology, and social-legal networks. *Curr. Bioinf.* **2011**, *6*, 94−121.
(134) Riera-Fernandez, P.; Munteanu, C. R.; Escobar, M.; Prado-Prado, F.; Martin-Romalde, R.; Pereira, D.; Villalba, K.; Duardo-Sanchez, A.; Gonzalez-Diaz, H. New Markov-Shannon entropy models to assess connectivity quality in complex networks: from molecular to cellular pathway, parasite-host, neural, industry, and legal-social networks. *J. Theor. Biol.* **2012**, *293*, 174−188.
(135) Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U. S. A.* **1982**, *79*, 2554−2558.
(136) Box, G. E. P.; Jenkins, G. M. *Time series analysis*; Holden-Day: San Francisco, 1970; p 553.
(137) Hill, T.; Lewicki, P. *STATISTICS Methods and Applications*; StatSoft: Tulsa, 2006.
(138) Mazurie, A.; Bonchev, D.; Schwikowski, B.; Buck, G. A. Evolution of metabolic network organization. *BMC Syst. Biol.* **2010**, *4*, 59.
(139) Kier, L. B.; Bonchev, D.; Buck, G. A. Modeling biochemical networks: a cellular-automata approach. *Chem. Biodivers.* **2005**, *2*, 233−243.
(140) Rosa da Silva, M.; Sun, J.; Ma, H. W.; He, F.; Zeng, A. P. Metabolic networks. In *Analysis of biological networks*; Junker, B. H., Schreiber, F., Eds.; Wiley & Sons: NJ, 2008, pp 233−253.
(141) Lee, D. S.; Park, J.; Kay, K. A.; Christakis, N. A.; Oltvai, Z. N.; Barabasi, A. L. The implications of human metabolic network topology for disease comorbidity. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 9880−9885.
(142) Rosenblatt, F. *Principles of neurodynamics; perceptrons and the theory of brain mechanisms*; Spartan Books: WA, 1962.
(143) Wasserman, S.; Faust, K. *Social network analysis: methods and applications*; Cambridge University Press: Cambridge, 1999.
(144) Fowler, J. H.; Jeon, S. The authority of Supreme Court precedent. *Social Networks* **2008**, *30*, 16−30.
(145) Duardo-Sánchez, A. Study of criminal law networks with Markov-probability centralities. In *Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks*; González-Díaz, H., Ed.; Bentham: Kerala, India, 2010; Chapter 12, pp 205−212.
(146) Duardo-Sánchez, A. Criminal law networks, markov chains, Shannon entropy and artificial neural networks. In *Complex Network Entropy: From Molecules to Biology, Parasitology, Technology, Social, Legal, and Neurosciences*; González-Díaz, H., Ed.; Bentham: Kerala, India, 2011; Chapter 8, pp 107−114.

**Artículo 3.**

**Predicción de Redes Complejas con Índices Markov-Balaban y ANNs**

MIANN Models of Networks of Bi ochemical Reactions, Ecosystems, and U.S. Supreme Court with Balaban-Markov Indices. A. Duardo -Sanchez, H. González-Díaz, A. Pazos
*Current Bioinformatics,* **2014***, 9, en imprenta*.

*Send Orders of Reprints at reprints@benthamscience.net*

***Current Bioinformatics*, 2014, *9*, 000 - 000**

# MIANN Models of Networks of Biochemical Reactions, Ecosystems, and U.S. Supreme Court with Balaban-Markov Indices

Aliuska Duardo-Sánchez [1,2,*], Humberto González-Díaz [3,4], and Alejandro Pazos[1]

[1] *Department of Information and Communication Technologies, University of A Coruña UDC, 15071, A Coruña, Spain.*

[2] *Department of Especial Public Law, Financial and Tributary Law Area, Faculty of Law, University of Santiago de Compostela, 15782, Santiago de Compostela, Spain*

[3] *Department of Organic Chemistry II, University of the Basque Country UPV/EHU, 48940, Leioa, Spain.*

[4] *IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain.*

**Abstract:** We can use Artificial Neural Networks (ANNs) and graph Topological Indices (TIs) to seek structure-property relationship. Balabans' J index is one of the classic TIs for chemo-informatics studies. We used here Markov chains to generalize the J index and apply it to bioinformatics, systems biology, and social sciences. We seek new ANN models to show the discrimination power of the new indices at node level in three proof-of-concept experiments. First, we calculated more than 1,000,000 values of the new Balaban-Markov centralities Jk(i) and other indices for all nodes in >100 complex networks. In the three experiments, we found new MIANN models with >80% of Specificity (Sp) and Sensitivity (Sn) in train and validation series for Metabolic Reactions of Networks (MRNs) for 42 organisms (bacteria, yeast, nematode and plants), 73 Biological Interaction Webs or Networks (BINs), and 43 sub-networks of U.S. Supreme court citations in different decades from 1791 to 2005. This work may open a new route for the application of TIs to unravel hidden structure-property relationships in complex bio-molecular, ecological, and social networks.

***Keywords*:** Artificial Neural Networks; Complex Networks; Metabolomics; Ecosystems; U.S. Supreme Court; Legal and Social Networks; Markov Chains.

## 1. INTRODUCTION

Complex networks are present in almost of levels of material world[1]. We see them after an inspection of complex networks formed at multi scales: from networks of metabolic organic reactions inside living beings[2], passing by interactions between species in food webs and ecosystems[3], until decisions in the U.S. Supreme Court[4, 5].

* **Corresponding author:** Duardo-Sánchez A., Email: aliuska.duardo@usc.es, *Department of Especial Public Law, University of Santiago de Compostela, 15782, Santiago de Compostela, Spain*

We make emphasis in the three previous examples because they will be subjected to closer inspection in this work. However, the number of examples is huge including also networks of Inorganic and Organic Reactions in the atmosphere of planets [6]; Obesity, or Alcoholism spreading in social networks[7, 8]; Internet and the World Wide Web (WW), Electric power distribution networks, *etc*. [9]. In any case, we can represent all these complex systems as a simple draw (graph) linking the parts of the system (nodes) by means of edges that express the existence of ties, interactions, or some relationship between nodes.

**1574-8936/14 $58.00+.00** 

## 2. REVIEW OF BALABANS' J INDEX

Network analysis may capture static or dynamic information about complex systems [10]. It is straightforward to realise, from the previous section, that we can use numeric parameters that quantify the different connectivity patterns in a graph to unravel information about hidden structure-property relationships (static or dynamic) in complex systems. Many authors call these numeric parameters as Connectivity or Topological indices (TIs) [11, 12].

TIs are numeric parameters used to describe the information about molecular structure of a molecule (molecular descriptor) or large systems (bio-molecular, social, technological, etc.) susceptible to be represented as a network or graph (G). For instance, we can define G ≡ (v, e) as a list of all pairs formed by the $v^{th}$ vertices or nodes (e.g.; atoms, proteins, countries...) interconnected by a list of $e^{th}$ edges, ties or relationships (chemical bonds, protein interactions, bilateral trade...). We can depict G in the plane (2D space) as a graphic representation of all this nodes (dots) interconnected by ties (edges). In any case, we can also construct graphs with vertices as points in a 3D space, and edges connecting these vertices. See for instance the work of Komosinski and Kubiak [13] in complexity about 3D morphologies.

Consequently, TIs are numerical parameters of a graph used to quantify information about the topology (in the sense of node connectivity) and are usually graph invariants (do not depend on node labels, rotation, translations, or geometric distortion of graph edges in general). In general, TIs are called to play an important role in the characterization of the complexity of real systems that can be represented as graphs or complex networks. For instance, Diudea *et al.*[14] published a paper in the journal complexity introducing a new super index based on Shell-matrices and polynomials.

Dehmer and Mowshowitz [15] published in the same journal a generalization of entropy measures applying information measures to a graph. In Prof. Balaban's words TIs are a convenient means of translating chemical constitution into numerical values[16].. Until now, multiple TIs have been defined and/or applied in some extend. Many of them are interrelated somehow and other TIs are more exotic. For instance, Basak and Balaban, *et al.*[17] calculated 202 molecular descriptors (TIs) for two chemical databases. The first contained 139 hydrocarbons and the second one a heterogeneous series of 1037 diverse compounds. They were able to group all these TIs into 14-18 clusters using variable cluster analysis. Correspondences between the same TIs in the two sets reveal how and why the various classes of TIs are mutually related and provide insight into what aspects of chemical structure they are expressing [17].

In any case, some of these TIs have become classic and/or TIs of reference with the pass of time. This is the case of Wiener[18-20], Hosoya[21, 22], Randic[23], Kier and Hall[24, 25], and other TIs. Taking into consideration the classification proposed by Balaban for the TIs[26], they can belongs to three generations or classes. The classes move from first-generation of TIs having the form of integer numbers with high degeneracy that limits their use to more elaborated TIs. In particular the same professor mentioned in the first sentence introduced one TIs; which have become one of the classics and is known as the Balaban's J index [16]. The J index is a function of q = number of edges in the molecular graph, μ = (q – n + 1) = the cyclomatic number of the molecular graph, n = number of atoms in the molecular graph. The J index also depends on $S_i$ = distance sums calculated as the sums over the rows or columns of the topological distance matrix **D** of the hydrogen-depleted graph G for the molecule. The formula of this classic TI is:

$$J(G) = \frac{q}{\mu + 1} \cdot \sum_{edges}^{q} \left( S_i \cdot S_j \right)^{-1/2} \qquad (1)$$

### Applications of in Molecular Informatics

Blaban's J index have been used in many chemo-informatics studies as input for many different Machine Learning (ML) algorithms; alone and/or combined with other TIs. Almost all applications of Balban's J index deal with drug discovery; in particular the prediction of drugs with higher biological activity and/or low toxicity. For instance, Sharma *et al.* [27]; synthesized new carboxylic acid ethyl esters and evaluated the *in vitro* antimicrobial and anticancer activity of these compounds. In a chemoinformatic study they demonstrated the importance of Balaban's J index and log P to describe the antimicrobial activity of these compounds.

Thakur *et al*. [28]; carried out a QSAR study on benzenesulphonamide carbonic anhydrase inhibitors using Balaban's J index. The regression analysis has shown that even in mono-parametric regression this index gave excellent results. Moreover, the combination of J with the first-order Randic connectivity index $^{1}\chi$ improved the results obtained. Krawczuk *et al.* [29]; used the J index and the electro-topological index for the prediction of retention data of polychlorinated biphenyls. J index has been used also as index of reference to test the performance of new TIs. In any case, the major number of applications of J index is in medicinal chemistry. Yadav, Kumar, and De Clercq [30] *et al.*, synthesized and evaluated the *in vitro* antimicrobial activity of new sulphonic acids. They compared the results of one-target *vs.* multi-target models for these compounds. The multi-target model based on Balaban index (J), a LUMO parameter, and second order valence connectivity index ($^{2}\chi_{v}$) was very useful to describe the antimicrobial activity of synthesized compounds. Naik, Dubey, and Kumar [31], development new predictive models for epipodophyllotoxin derivatives. Epipodophyllotoxins are important anticancer drugs used in chemotherapy for various types of cancers. The model was obtained with descriptors such as solvent-accessible surface area, heat of formation, Balaban index, number of atom classes, and sum of E-state index of atoms. Fernandes [32] predicted the activity of forty-three pyrazinoates against *M. tuberculosis* ATCC 27294, using Balaban index (J) and other parameters as inputs for a genetic a algorithm function and partial least squares regression (WOLF 5.5 program).

Panaye, Doucet, and Devillers, *et al.* [33]; developed compared decision trees *vs.* support vector machine (SVM) for classification of androgen receptor ligands. They predicted relative binding affinity (RBA) to a large set of about 200 chemicals with descriptors calculated from CODESSA software including hydrophobicity parameter (logP), Balaban index, and other descriptors. Zavrsnik, Muratović, and Spirtović[34]; reported the synthesis of 4-arylaminocoumarin derivatives with antimycotic effects and a QSAR study of this activity with, Balaban J index, Wiener W index, and other TIs and physicochemical properties.

Ma, Chen, and Yang[35]; used the Balaban index and other molecular descriptors to model blood-brain barrier (BBB) penetration of different compounds. One aspect of special relevance for this work is the use of Balaban's J index as input for Artificial Neural Networks (ANNs). Dashtbozorgi and Golmohammadi [36], used Multiple Linear Regression (MLR) and ANNs to model water-to-wet butyl acetate partition coefficient of 76 organic solutes using Balaban index (J) together with the Kier and Hall index of order 2 ($^{2}\chi$), and other indices. Jalali-Heravi and Fatemi [37]; developed ANN model using Balaban index and other parameters to predict the thermal conductivity detection response factors of 110 organic compounds.

In other line of thinking, some author have reported applications of Balban's J index in mathematical and/or physical-chemistry; including the generalization of this index to create other TIs (called Balaban type parameters). Dehmer *et al.* [38] compared the discriminative power for graphs of a new information index *vs.* the J values for benchmark dataset of nearly 12 million exhaustively generated, non-isomorphic, and unweighted graphs. In this sense, Basak *et al.*[39]; reviewed the use of mathematical structural invariants in analyzing combinatorial libraries.

Ratkiewicz and Truong[40]; reported a new method for automatic generation of mechanisms of reactions of complex systems using Balaban, Schulz, Connectivity and other TIs. Rastija and Medić-Sarić [41]; studied antioxidant activity of wine polyphenols with Balaban index, Balaban-type index and other descriptors. Randić and Pompe[42]; reported the variable Balaban J index and the "reversed" Balaban index 1/J as well as a novel index 1/JJ derived from J and 1/J.

## 3. MIANN MODELS

Dehmer *et al.* [43]; noted that many TIs of complex networks have been developed without giving a proper proof of their potential applications. They also talk about the high interest in the development of software packages to calculate TIs. We can use computer software like CENTIBIN [44], PAJEK [45], and QUACN [43, 46] can be used to calculate TIs of complex networks.

Specifically, different authors have reported, and/or used different types of TIs that may be considered variations or generalizations of the J index in some extend. For instance, Randić & Pompe [42]; introduced several variable molecular descriptors, derived from the distance matrix and the "reversed" distance matrix. This includes the variable Balaban J index and the "reversed" Balaban index 1/J as well as a novel index 1/JJ derived from J and 1/J. Balaban type indices: J, Jz, Jm, Jv, Jc, and Jp have been used to predict the supra-molecular complexing ability of a sulfonamides[47].

González-Díaz *et al.* [48]; developed the software MARCH-INSIDE. The software MI uses Markov Chains theory to calculate TIs of order k ($TI_k$) inside very different structures in many file formats such as SMILE .txt and .mol file of drugs, .pdb files of proteins, or .ct files for RNAs. In any case, MI cannot upload formats of Complex Networks (.mat, .net, .dat, *etc*.). Accordingly, we have inspired in MI to write a new software called MI-NODES (MARCH-INSIDE NOde DEScriptors). MI-NODES can upload many formats of complex networks and calculate several $TI_k$ values for these networks. MI-NODES incorporates node-node transition Markov probabilities ($p_{ij}$) inside the formula of classical TIs. In so doing, we can calculate new versions of classic TI(G) of a full graph G based on Markov Chains $TI_k(G)$ and the respective node centralities $TI_k(i)$ [49-52].

We can combine MI with different Machine Learning algorithms. For example, we can combine MI with Linear Discriminant Analysis (LDA) implemented in STATISCA [53] or ML methods implemented in WEKA [54]. In particular, we can combine MI with Artificial Neural Networks (ANNs); which are very powerful bio-inspired Artificial Intelligence (AI) algorithms. The name of this strategy is MIANN (MI & ANN models). We have reviewed the MIANN in a recent work. In this work we focused on theoretical basis, development of web servers, and applications in molecular sciences [55]. In any case, almost al MIANN models use the original MI software and focus in chemical or bio-molecular systems. Conversely, the use of MI-NODES to seek MIANN models is relatively new direction.

Taking into consideration all the aspect discussed in the introduction we decided to perform the following work. In this work we generalized the Balaban's J index in a different direction, implement the calculation of these TIs in MI-NODES, and seek new MIANN models to show in proof-of-concept experiments to demonstrate the discrimination power of the new indices.

## Theory of MIANN models

In many complex networks we can found certain uncertainty on the assignation of links between nodes due to errors or existence of contradictory information obtained by different experimental ways. Let be the connectivity pattern $\mathbf{L}_i$ (set of all walks of length k from the $i^{th}$ node to the remnant *n* -1 nodes in the network) we can try to seek a model to discriminate the observed (correct) $\mathbf{L}_i$ of all nodes in real networks *vs.* not correct connectivity patterns. The output of this type of models may be a real value function $S_j$ used to score the quality of a given connectivity pattern $\mathbf{L}_i$.

In this sense, according to the model, the higher is the value of $S_j$ the closer to the correct pattern are the links set for $j^{th}$ in the network as a whole. On the other hand, the input dependent variable $L_i = 1$ when a node is correctly linked to the rest of nodes (real or correct $L_i$) in the network or $L_i = 0$ for nodes that we know *a priori* that present an incorrect connectivity pattern. We can use different TIs and/or node centralities to describe numerically different connectivity patterns **L**i (correct or incorrect patterns in a give network). Next, we can use these TIs as inputs of ANN algorithms in order to search for non-linear and/or linear models able to predict the correct **L**is. In the particular case of a linear MIANN equation obtained by LNN we can write the general formula of this type of MIANN models as follows:

$$S_j = \sum_{k=0}^{5} a_k \cdot TI_k(j) + \sum_{g=0}^{g=Ng} \sum_{k=0}^{5} b_{gk} \cdot \left[ TI_k(j) - TI_k(j)_{g-avg} \right] + c_0 \quad (2)$$

$$= \sum_{k=0}^{5} a_k \cdot TI_k(j) + \sum_{g=0}^{g=Ng} \sum_{k=0}^{5} b_{gk} \cdot \Delta TI_k(j)_g + c_0$$

In these equations TIk are node centralities based on topological indices calculated with software MI-NODES based on the MARCH-INSIDE algorithm. However, we can use any TI calculated with other software. The coefficients ($a_k$) quantify the influence of the centralities of nodes used as input. The coefficients ($^g b_k$) quantify the effect of the deviation of the TIs of a given node with respect to

the average value of the TIs of a sub-set of nodes that obey a given condition in a network of reference correctly constructed. The deviation terms have the general form $\Delta TI_k(j)_g = [TI_k(j) - TI_k(j)_{g.avg}]$. Where, $TI_k(j)_{g.avg}$ is the average value (avg) of $TI_k(j)$ for a sub-set or group (g) of nodes of the same graph G ($g \in G$) that obey a given condition.

This type of deviation terms resembles the moving average terms used in time series models like in Box–Jenkins' ARIMA models [56]. However, in the present work g may be not only a time frame or season (laws approved in the same year) but also a biological boundary (metabolic reactions in the same organism) or spatial condition (interactions in the same eco-system); see results section. Specifically, we can write the linear the equation of the MIANN model obtained by LNN analysis for MRNs as follows:

$$S_i = \sum_{k=0}^{5} a_k \cdot TI_k(i) + \sum_{k=0}^{5} b_{gk} \cdot \left[TI_k(i) - TI_k(i)_{Org.avg}\right] + c_0 \quad (3)$$

$$= \sum_{k=0}^{5} a_k \cdot TI_k(i) + \sum_{g=0}^{g=Ng} \sum_{k=0}^{5} b_{gk} \cdot \Delta TI_k(i)_{Org} + c_0$$

The LNN model for the particular case of BINs of the IWDB has the following formula:

$$S_i = \sum_{k=0}^{5} a_k \cdot TI_k(i) + \sum_{k=0}^{5} b_{gk} \cdot \left[TI_k(i) - TI_k(i)_{Web.avg}\right] + c_0 \quad (4)$$

$$= \sum_{k=0}^{5} a_k \cdot TI_k(i) + \sum_{k=0}^{5} b_{gk} \cdot \Delta TI_k(i)_{Web} + c_0$$

Last, the LNN model for the particular case of USSCN has the following formula:

$$S_i = \sum_{k=0}^{5} a_k \cdot TI_k(i) + \sum_{k=0}^{5} {}^1b_k \cdot \left[TI_k(i) - TI_k(i)_{Year.avg}\right]$$

$$+ \sum_{k=0}^{5} {}^2b_k \cdot \left[TI_k(j) - TI_k(j)_{Citing avg}\right] + c_0 \quad (5)$$

$$= \sum_{k=0}^{5} a_k \cdot TI_k(i) + \sum_{k=0}^{5} {}^1b_{gk} \cdot \Delta TI_k(i)_{Year}$$

$$+ \sum_{k=0}^{5} {}^2b_{gk} \cdot \Delta TI_k(i)_{Citing} + c_0$$

Where, the $J_k(j)$ values are the Balaban-Markov centrality parameters of a given j-th judicial cases resolved by the U.S. Supreme Court. Whereas, the $J_k(j)_{Year.avg}$ and $J_k(j)_{Citing.avg}$ values are the average of these parameters for all the cases in this court in the given year (Year.avg) or for the cases citing $j^{th}$ (Citing.avg). In addition, we can use statistical parameters of both train and external validation series to assess the goodness-of-fit of ANN models: *n* = number of cases, Specificity, and Sensitivity[57].

## 4. NEW MIANN MODELS WITH $J_k$ INDICES

In the present work, we introduce by the first time a new type of Balaban type indices called Balaban-Markov. The global Balaban-Markov $J_k(G)$ indices are useful to study the full graph G of a molecular or complex network system in general. The respective node centralities $J_k(i)$ are expected to be useful in the study complex networks at node level. The calculation of the new indices is implemented in MI-NODES. Subsequently, we applied the MIANN strategy to study complex bio-molecular complex networks. In addition, we studied ecological and social-legal systems as well to illustrate the uses of the new parameter in higher scales. In closing, we developed three proof-of-concept experiments to illustrate the potentialities of the new parameters. In each experiment, we seek new MIANN models by the first time useful to discriminate correct from unreliable connectivity patterns in complex networks. In the first experiment, we found new MIANN-Balaban models for Metabolic Reactions of Networks (MRNs) for 43 organisms (bacteria, yeast, nematode and plants). In the second experiment, we developed MIANN-Balaban for 73 Biological Interaction Webs or Networks (BINs). The biological interactions or relationships present in these networks include: prey-hunter, parasite-host, plant-seed disperser, anemone-clown fish species and others. In the third experiment, we found MIANN-Balaban models for >40 networks relevant for Legal sciences. Each one of these networks (5KCNs) represent one slot of 5000 (5K) cites to 1000-5000 cases of the U.S. Supreme court in different decades (1700-2006). Despite the differences between the different networks the workflow used in all the experiments is essentially the same (see **Figure 1)**. It is straightforward to realise from **Figure 1** that the steps given are essentially the same used in QSAR/QSPR analysis of molecular systems in chemoinformatics. The datasets used to develop the new MIANN models are the following.

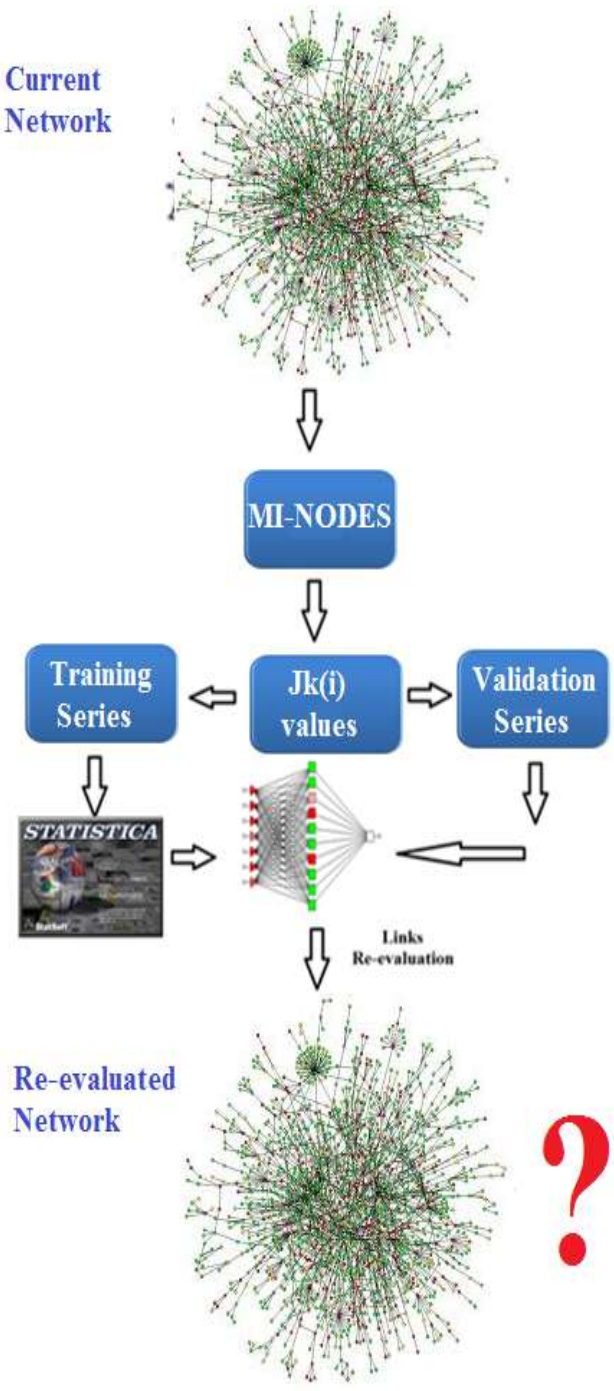



**Figure 1.** General workflow used in this work

**Data 1: Metabolic Reaction Networks (MRNs)**

These MRNs were downloaded in a zipped ASCII file directly from Barabasi's group web (http://www.nd.edu/~networks/resources.htm). In this file each number represents a substrate in the metabolic network of corresponding organism. Data-format is: From → To (directed link). The information studied was previously obtained by Jeong *et al*. from the 'intermediate metabolism and bioenergetics' portions of the WIT database and used in order to try to understand the large-scale organization of metabolic networks [2]. According to the authors, biochemical reactions described within a WIT database are composed of substrates and enzymes connected by directed links. For each reaction, educts and products were considered as nodes connected to the temporary educt-educt complexes and associated enzymes. Bidirectional reactions were considered separately. For a given organism with N substrates, E enzymes and R intermediate complexes the full stoichiometric interactions were compiled into an (N+E+R) X (N+E+R) matrix, generated separately for each of the different organisms. In **Table 1** we report the values of the $J_k(G)$ indices for the MRNs of different organisms studied in this work. We also depict the values of classic parameters of these MRNs that have been reported by other authors. In the next sections we shall discuss the relationships between $J_k(G)$ indices and these classic parameters.

**Data 2: Biological Interaction Networks (BINs)**

In a recent review we discussed many biological interaction webs or networks (BINs) including those contained in the IWDB [58]. In a previous work, we downloaded all matrices compiled in the IWDB and transformed them into BINs in .net format. This format list all pairs (arcs or edges) of species (nodes) into a text file [59]. The IWDB (http://www.nceas.ucsb.edu/interactionweb/resources.html) currently contains datasets on species interactions from several communities in different parts of the world. Data included many types of ecological interactions: plant-pollinator, plant-frugivore, plant-herbivore, plant-ant mutualism and predator-prey interactions. Most webs are "bipartite networks", which consist of two groups that are assumed to interact with species in the other group but not with species in their own group (*e.g.*, plants and insect herbivores). In **Table 2** we can find a summary of all the available datasets; see a list of references to original sources in our previous work [59]. The full list of reference is too large to be cited here.

**Table 1.** Average values $J_k(i)_{org.avg}$ *vs.* classic parameters of MRNs of some organisms

| MRN | $J_0$ | $J_1$ | $J_2$ | $J_3$ | $J_4$ | $J_5$ | N | $L_{in}$ | $L_{out}$ | R | E | $g_{in}$ | $g_{out}$ | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AA | 3289.9 | 3233.4 | 1863.0 | 1117.6 | 774.8 | 677.8 | 419 | 1278 | 1249 | 401 | 285 | 2.10 | 2.20 | 3.30 |
| AB | 2933.9 | 2869.2 | 1625.4 | 975.6 | 682.6 | 634.4 | 395 | 1202 | 1166 | 380 | 271 | 2.10 | 2.20 | 3.20 |
| AG | 4234.2 | 4165.9 | 2432.1 | 1471.4 | 1028.9 | 910.0 | 496 | 1527 | 1484 | 486 | 299 | 2.20 | 2.20 | 3.50 |
| AP | 1514.3 | 1474.5 | 806.4 | 458.4 | 315.1 | 328.9 | 204 | 588 | 575 | 178 | 135 | 2.20 | 2.20 | 3.20 |
| AT | 2324.1 | 2284.3 | 1343.7 | 822.8 | 580.1 | 508.9 | 302 | 804 | 789 | 250 | 185 | 2.10 | 2.30 | 3.50 |
| BB | 1085.8 | 1057.7 | 581.7 | 336.6 | 235.8 | 225.1 | 187 | 442 | 438 | 140 | 106 | 2.30 | 2.40 | 3.00 |
| BS | 6513.1 | 6268.7 | 3314.5 | 1957.0 | 1331.8 | 1204.5 | 785 | 2794 | 2741 | 916 | 516 | 2.20 | 2.10 | 3.30 |
| CA | 4018.9 | 3938.0 | 2220.9 | 1313.5 | 904.5 | 794.6 | 494 | 1624 | 1578 | 511 | 344 | 2.10 | 2.20 | 3.30 |
| CE | 3707.9 | 3642.9 | 2067.0 | 1224.1 | 842.8 | 757.7 | 462 | 1446 | 1418 | 450 | 295 | 2.10 | 2.20 | 3.30 |
| CJ | 2834.7 | 2774.5 | 1595.4 | 959.9 | 668.4 | 584.8 | 380 | 1142 | 1115 | 359 | 254 | 2.10 | 2.30 | 3.20 |
| CL | 2821.7 | 2742.7 | 1479.3 | 839.2 | 561.1 | 556.9 | 389 | 1097 | 1062 | 333 | 231 | 2.10 | 2.20 | 3.30 |
| CQ | 1238.6 | 1193.7 | 630.6 | 352.4 | 242.9 | 253.5 | 194 | 401 | 391 | 134 | 84 | 2.20 | 2.30 | 3.40 |
| CT | 1491.2 | 1466.7 | 904.3 | 576.3 | 419.2 | 374.0 | 215 | 479 | 462 | 158 | 94 | 2.20 | 2.40 | 3.50 |
| CY | 4234.7 | 4089.9 | 2101.2 | 1181.1 | 788.0 | 804.8 | 546 | 1782 | 1746 | 570 | 370 | 2.00 | 2.20 | 3.30 |
| DR | 6805.0 | 6614.5 | 3609.3 | 2147.1 | 1480.5 | 1345.6 | 815 | 2870 | 2811 | 965 | 557 | 2.20 | 2.10 | 3.30 |
| EC | 6596.4 | 6445.0 | 3629.6 | 2185.4 | 1509.8 | 1363.7 | 778 | 2904 | 2859 | 968 | 570 | 2.20 | 2.10 | 3.20 |
| EF | 2987.0 | 2936.2 | 1699.0 | 1010.1 | 693.0 | 608.0 | 386 | 1244 | 1218 | 382 | 281 | 2.10 | 2.20 | 3.10 |
| EN | 2870.5 | 2814.5 | 1585.1 | 937.1 | 645.8 | 576.2 | 383 | 1095 | 1081 | 339 | 254 | 2.10 | 2.20 | 3.30 |
| HI | 4161.9 | 4019.0 | 2098.8 | 1239.7 | 863.2 | 855.3 | 526 | 1773 | 1746 | 597 | 361 | 2.10 | 2.30 | 3.20 |
| HP | 2833.8 | 2759.3 | 1486.2 | 867.4 | 594.0 | 547.4 | 375 | 1181 | 1144 | 375 | 246 | 2.00 | 2.30 | 3.30 |
| MB | 3228.1 | 3166.7 | 1835.6 | 1112.3 | 779.1 | 683.3 | 429 | 1247 | 1221 | 391 | 282 | 2.20 | 2.20 | 3.20 |
| MG | 1529.8 | 1498.1 | 926.5 | 607.2 | 457.3 | 413.8 | 209 | 535 | 525 | 196 | 85 | 2.40 | 2.20 | 3.50 |
| MJ | 3587.0 | 3523.2 | 2049.2 | 1250.6 | 880.8 | 773.3 | 424 | 1317 | 1272 | 415 | 264 | 2.20 | 2.30 | 3.50 |
| ML | 3254.8 | 3199.0 | 1881.4 | 1166.1 | 835.1 | 744.5 | 422 | 1271 | 1244 | 402 | 282 | 2.20 | 2.20 | 3.20 |
| MP | 1287.6 | 1260.7 | 731.9 | 427.0 | 293.8 | 266.7 | 178 | 470 | 466 | 154 | 88 | 2.30 | 2.20 | 3.20 |
| MT | 4564.6 | 4412.8 | 2317.0 | 1288.3 | 850.5 | 871.8 | 587 | 1862 | 1823 | 589 | 358 | 2.00 | 2.20 | 3.30 |
| NG | 3168.0 | 3115.1 | 1824.2 | 1107.5 | 771.0 | 663.6 | 406 | 1298 | 1270 | 413 | 285 | 2.10 | 2.20 | 3.20 |
| NM | 2905.4 | 2851.4 | 1612.5 | 958.2 | 662.7 | 577.5 | 381 | 1212 | 1181 | 380 | 271 | 2.20 | 2.20 | 3.20 |
| OS | 2289.5 | 2252.1 | 1309.9 | 784.8 | 544.9 | 483.8 | 292 | 763 | 751 | 238 | 178 | 2.10 | 2.30 | 3.50 |
| PA | 5751.4 | 5532.1 | 2915.7 | 1667.5 | 1105.3 | 1079.6 | 734 | 2453 | 2398 | 799 | 490 | 2.10 | 2.20 | 3.30 |
| PF | 2453.1 | 2409.0 | 1414.0 | 846.6 | 580.0 | 498.9 | 316 | 901 | 867 | 283 | 191 | 2.00 | 2.30 | 3.40 |
| PG | 3203.6 | 3146.9 | 1875.8 | 1154.1 | 813.9 | 711.5 | 424 | 1192 | 1156 | 374 | 254 | 2.20 | 2.20 | 3.30 |
| PH | 2491.1 | 2448.1 | 1448.4 | 875.7 | 604.3 | 526.1 | 323 | 914 | 882 | 288 | 196 | 2.00 | 2.20 | 3.40 |
| PN | 3211.2 | 3146.1 | 1784.3 | 1068.3 | 743.5 | 651.5 | 416 | 1331 | 1298 | 412 | 288 | 2.10 | 2.20 | 3.20 |
| RC | 5344.0 | 5163.4 | 2697.5 | 1526.3 | 1016.9 | 1098.9 | 670 | 2174 | 2122 | 711 | 427 | 2.10 | 2.20 | 3.40 |
| RP | 1466.1 | 1437.9 | 818.8 | 476.1 | 329.3 | 312.3 | 214 | 510 | 504 | 155 | 100 | 2.30 | 2.30 | 3.40 |
| SC | 4591.4 | 4455.4 | 2294.2 | 1285.9 | 859.3 | 835.6 | 561 | 1934 | 1889 | 596 | 402 | 2.00 | 2.20 | 3.30 |
| ST | 3071.1 | 2998.9 | 1650.3 | 984.3 | 683.9 | 636.4 | 403 | 1300 | 1277 | 404 | 280 | 2.10 | 2.20 | 3.10 |
| TH | 3624.3 | 3554.8 | 2018.7 | 1203.8 | 834.9 | 742.0 | 430 | 1374 | 1331 | 428 | 280 | 2.20 | 2.20 | 3.40 |
| TM | 2448.2 | 2396.7 | 1338.1 | 770.7 | 522.1 | 471.6 | 338 | 1004 | 976 | 302 | 223 | 2.10 | 2.20 | 3.20 |
| TP | 1331.5 | 1300.7 | 715.5 | 414.4 | 285.7 | 274.8 | 207 | 562 | 555 | 175 | 124 | 2.20 | 2.30 | 3.10 |

**Data 3: U.S. Supreme Court Networks**

We used a complex network constructed by Fowler *et al.*[4]. The authors included 26,681 majority opinions written by the U.S. Supreme Court. The dataset contains all cases that cite this U.S. Supreme Court decisions from 1791 to 2005. In this network each case is represented by a node. The links between two nodes $a_i$ and $a_j$ (arcs) express that the case $j^{th}$ cites the $i^{th}$ case previous to it (precedent). In order to carry out a study more focused on specific intervals of time and also study the effect of accumulation of cites we split the data set of sub-sets (sub-networks). Each one of these sub-networks represent one slot of 5000 (5K) cases that cites 1000-5000 cases of the U.S. Supreme court in different decades. We call these networks as the 5K-Citations Network (5KCNs). It also make more tractable the dataset for computation of $J_k(i)$ values. We constructed in total 43 5KCNs involving approximately 5000.43 > 22,000 cases of the U.S. Supreme court.

## 5. DISCUSSION

**New Balaban-Markov Centralities for nodes**

As we mention in the introduction, Dehmer *et al.* [43]; discussed this problem and suggested the classification of complex biological networks as a proof-of-concept experiment in this sense. They also concluded that there is a high necessity for freely available software packages to calculate and validate new TIs of complex networks. Different authors have reported, and/or used in QSAR studies, different types of TIs that may be considered variations or generalizations of the J index in some extend. For instance, Randić & Pompe [42]; introduced several variable molecular descriptors, derived from the distance matrix and the "reversed" distance matrix. This includes the variable Balaban J index and the "reversed" Balaban index 1/J as well as a novel index 1/JJ derived from J and 1/J. Balaban type indices: J, Jz, Jm, Jv, Jc, and Jp have been used to predict the supra-molecular complexing ability of a sulphonamides [47].

In this work we generalized the Balaban's J index in a different direction, implement it in a new software (MI-NODES), and carry out proof-of-concept experiments to demonstrate the discrimination power of the new indices. The product of the parameters $S_i \cdot S_j$ plays a central role in the definition of J (see eq. 1). In the introduction, we remember to the reader that $S_i$ and $S_j$ are distance sums calculated as the sums over the rows or columns of the topological distance matrix **D**. As a consequence, the result of unique for a given graph G and we obtain only one value of J. However, we can weight this product with the probabilities ${}^{k}p_{ij}$ with which both nodes are connected by walks of length k. In this case, we can generalize the J index to a series of higher order analogues $J_k$ that quantify the probability of interconnection of all pairs of nodes at different orders. The values of ${}^{k}p_{ij}$ are the elements of the Markov matrix (${}^{1}\mathbf{\Pi}$) used in the MARCH-INSIDE algorithm. We can construct it as follows: first, we download from public resources the connectivity matrix **L** or obtain the data about the links between the nodes to assemble **L** (*n* by *n* matrix, where *n* is the number of vertices). Next, the Markov matrix **Π** is built. It contains the vertices probability ($p_{ij}$) based on **L**. The probability matrix is raised to the power *k*, resulting $({}^{1}\mathbf{\Pi})^{k}$. The resulting matrices ${}^{k}\mathbf{\Pi}$ are the $k^{th}$ natural powers of ${}^{1}\mathbf{\Pi}$ and contain the transition probabilities ${}^{k}p_{ij}$. These are the probabilities to reach the $j^{th}$ node moving from the $i^{th}$ node throughout a walk of length *k* (for each *k*). Also other authors, like Estrada [60], has used the same type of matrix to generate indices of complex networks more recently. In this work, we report by the first time this new type of $J_k$ index. We also give the definition of the node centralities $J_k(i)$ for nodes of complex networks based on the same idea. In this case we can obtain a total of *k* values of new Balaban-Markov indices $J_k(G)$ and centralities $J_k(i)$ for each graph or each $i^{th}$ node, using the following formula:

$$J_k(G) = \frac{q}{\mu + 1} \cdot \sum_{edges}^{q} \left( {}^{k}p_{ij} \cdot S_i \cdot S_j \right)^{-1/2} \quad (6)$$

$$J_k(i) = \frac{q}{\mu + 1} \cdot \sum_{i-edges}^{z_i} \left( {}^{k}p_{ij} \cdot S_i \cdot S_j \right)^{-1/2} \quad (7)$$

**Table 2.** Summary of some BINs included in the IWDB.

| BIN | Habitat type | Location | #OA[a] | #OB[a] | BIN | Habitat type | Location | #OA[a] | #OB[a] |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Coral reefs | Pacific | 10 | 26 | 21 | Arctic community | Canada | 29 | 86 |
| 2 | Freshwater lake | Canada | 7 | 29 | 22 | Heathland habitat | Mauritius Is. | 135 | 74 |
| 3 | Freshwater lake | Canada | 10 | 40 | 23 | Beech forest | Japan | 93 | 679 |
| 4 | Freshwater lake | Canada | 31 | 144 | 24 | High Arctic | Canada | 32 | 115 |
| 5 | River | Canada | 14 | 51 | 25 | Montane forest | Australia | 42 | 91 |
| 6 | River | Canada | 17 | 53 | 26 | Multiple communities | Galapagos Is. | 106 | 54 |
| 7 | Freshwater lake | Canada | 33 | 97 | 27 | Xeric scrub | Argentina | 21 | 45 |
| 8 | Freshwater reservoir | Canada | 6 | 25 | 28 | Medow | UK | 25 | 79 |
| 9 | rainforest | Australia | 51 | 41 | 29 | Arctic community | Canada | 11 | 18 |
| 10 | rainforest | Peru | 8 | 18 | 30 | Deciduous forest | USA | 13 | 44 |
| 11 | tropical forest | C. Rica | 6 | 4 | 31 | Coastal forest | Mauritius Is. | 14 | 13 |
| 12 | Amazon rainforest | Brazil | 16 | 25 | 32 | Upland grassland | S. Africa | 9 | 56 |
| 13 | Arid grasslands | USA | 54 | 24 | 33 | Palm swamp | Venezuela | 33 | 53 |
| 14 | Whole country | Finland | 5 | 64 | 34 | Agricultural area | USA | 456 | 1429 |
| 15 | Andean scrub | Chile | 87 | 98 | 35 | Caatinga | Brazil | 51 | 25 |
| 16 | Boreal forest | Canada | 12 | 102 | 36 | Maple-oak woodland | USA | 7 | 32 |
| 17 | Caatinga | Brazil | 13 | 13 | 37 | Peat bog | Canada | 13 | 34 |
| 18 | Mt. forest and grassland | USA | 96 | 276 | 38 | Montane forest | Argentina | 10 | 29 |
| 19 | High-altitude desert | Canary Is. | 11 | 38 | 39 | Forest | Papua | 31 | 9 |
| 20 | Alpine subarctic comm. | Sweden | 23 | 118 | 40 | Tropical forest | Panama | 13 | 11 |

**MIANN models of MRNs**

With the development of systems biology the study of Metabolic Reaction Networks (MRNs) is gaining in importance due to possible applications in Biotechnology[61, 62] and Biomedicine with the study of disease comorbidity[63] has been approached with network topology methods. In **Table 1** we can find a summary of the properties of the MRNs studied. In this work, we used the $J_k(j)$ values and other parameters of MRNs presented in **Table 1** to carry out a Two-Way Joining cluster analysis (TWJCA) of this dataset; see **Figure 2 (A)**.

We used a Number of variables: 26 and Number of cases = 43 and found a Number of blocks = 193 with a Threshold computed from data = 0.4941518 (StDv/2), Total Sample Mean = -0.00, and Standard Deviation = 0.9883036. The results of this TWJCA are important to unravel hidden relationships between the new $J_k(j)$ indices with classic TIs as well as other parameters of the MRNs for each one of the 43 organisms studied. TWJCA shows that $J_k(j)$ indices form a "weak" cluster with some TIs for a group of networks but the TWJCA do not detect clear clusters. In general, TWJCA shows that $J_k(j)$ values seems to codify useful structural information of MRNs that it is not trivially related to the information codified by other parameters because there are not strong clusters formed between them.

Jeong *et al*. [2], noted that known MRNs models of different organisms present similar topological properties and not all metabolic pathways have been confirmed experimentally, and the experimental corroboration of a metabolome of one organism is a very hard task. Consequently, we need alignment-free techniques to evaluate the correct connectivity patterns $L_i$ for all nodes in MRNs. Here we developed different MIANN models based on $J_k(i)$ values to predict correct connectivity patterns Li of nodes in MRNs of a large number of organisms (43

in total). The best MIANN model found was the a non-linear three layers perceptron MLP 9:9-6-1:1. It has very good values of Accuracy (Ac), Sensitivity (Sn), and Specificity (Sp) > 81% in both learning and external validation series. By the contrary, the best linear model found do not shows very good results with Ac, Sn, and Sp < 65% in all cases. This result indicates that $J_k(i)$ values are very powerful descriptors. Note that with a non-linear but relatively simple MLP (only 9 inputs and 6 hidden neurons) we can predict in 43 organisms a very high number of 29117 metabolic reactions in learning and 9729 in validation. In **Table 3** we depict the classification matrices and the topology of the MIANN models discussed.

**MIANN models of BINs**

In analogy to the study of MRNs, we used the $J_k(j)$ values of BINs for different ecosystems and/or food webs. In **Table 2** (see previous section), we showed the names, location, number of species, and reference for many of these networks in order to illustrate the high complexity of the dataset studied. Almost all networks studied are directed and also bipartite in some cases. It means that the biological interaction is in almost all cases between species of a group (A) with a given biological function with species of a second group with another function in many cases complementary to A somehow. For instance, we can find anemones, plants, or predators in the first group and parasite, herbivore, pollinator, prey, or seed disperser species in the second group. Anyhow, there also many complex networks more complicated situations typical of an imbricate food web were many species may interact with the first or the second function in different cases. For instance, prey specie 1 is predated by predator 2, which is in turn predated / hunted by specie 3. Similarly, may appear hyperparasitism relationships were host specie 1 is parasited by specie 2 that is in turn the host (is parasited by) of specie 3.

Here, we used also TWJCA to try to unravel relationships between $J_k(j)$ values (of BINs in this case) with other parameters of complex networks. In this second TWJCA experiment we included classic TIs calculated with MI-NODES.

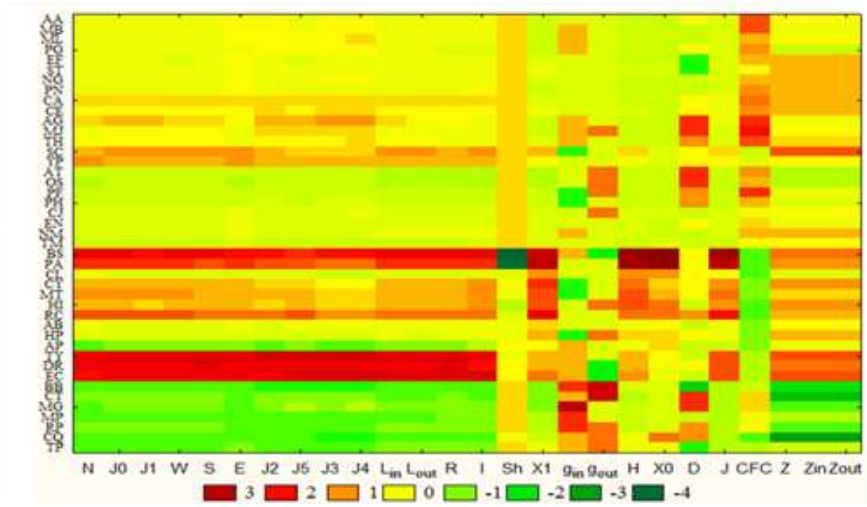
Some of these indices are the Wiener index (W), Shannon entropy (Sh), Gutman topological index (I), Schultz index (S) as well as $\chi$ connectivity indices of Randic or Kier & Hall. We also calculated with MI-NODES some node centralities of BINs like total, in, and out node degrees (Z, $Z_{in}$, $Z_{out}$) as well as Current Flow Centrality (CFC), and Current Flow Betweeness (CFB). TWJCA also pointed to $J_k(j)$ values as useful TIs that codify new structural information of BINs as they are not strongly clustered with other classic TIs. In **Table 4** we depict the $J_k(j)$ values of many BINs. We also give $Z_{in}$ and $Z_{out}$ values for comparative purposes. We also added Wiener W values (proportional to the sum of all topological distance between the specie and all other species in the web. In **Figure 2 (B)** we illustrate the results of the TWJCA experiment.

We also tested different MIANN models with linear and non-linear topology. Classification results in training and validation series for both MIANN models appear in **Table 5**. Linear MIANN models of BINs present slightly better results with respect to MRNs. The Ac, Sn, and Sp values here are close to 70% in all cases. However, once again the best model found is a MLP. The model MLP 12:12-9-1:1 present Ac, Sp, and Sn values higher than 81%. The model is similar in goodness-of-fit and topology to the best MIANN model for MRNs.

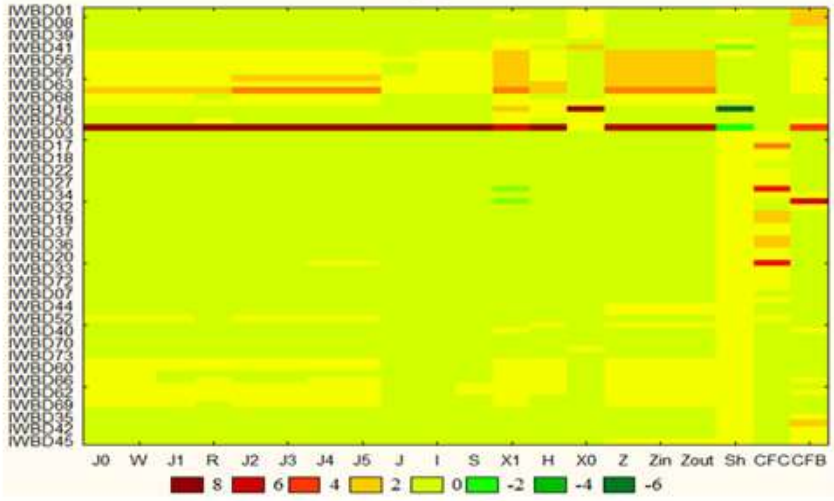
**MIANN models of U.S. Supreme Court**

There is a long tradition on the application of complex networks methods in social sciences; known as social network analysis (SNA) since 1930[64]. Using SNA we can unravel non-linear relationships between different laws and try to predict for instance the effect of these laws in society.

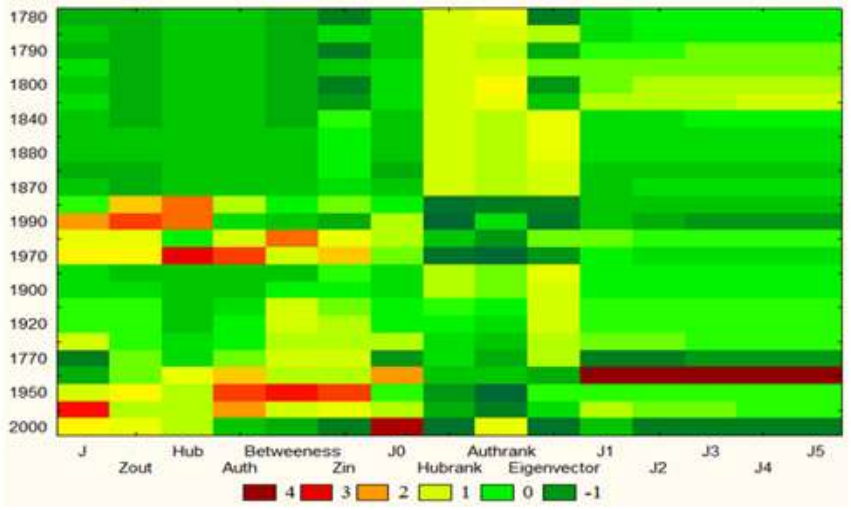
A turning point in this direction is the network constructed by Fowler *et al*. [4]. It represents a wonderful source (possibly the more complete) for the study of dynamics (changes along time) in the U.S. Supreme Court. The authors have withdrawn all cases cited in the text of each majority opinions from 1754 to 2002. According to them, opinion writers may cite a case just to mention it in passing or because they disagree. Legal rules are cited to provide legal justifications even if it is not a reliance on authority. Thus, they included all judicial citations in the dataset (including various types of citations) that could link cases together.




**(A)**



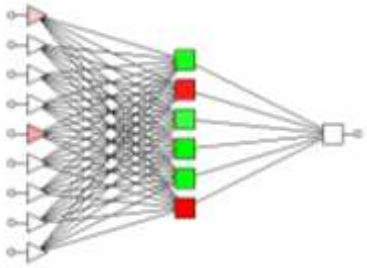
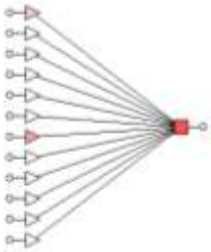

**(B)**




**(C)**

**Figure 2.** Cluster analysis of $J_k(j)$ values *vs.* classic TIs and other parameters of: (A) MRNs of 43 organisms, (B) BINs of 73 ecosystems, and (C) 73 US Supreme Court 5KCNs

**Table 3.** Linear vs. Non-linear MIANN models of MRNs of 43 organisms based on $J_k$ centralities

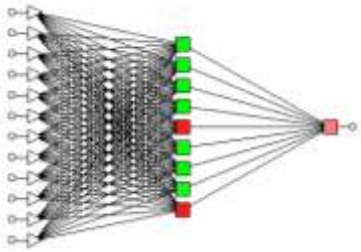
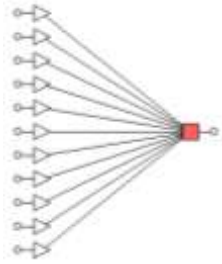
| MIANN models | Li | Li = 1 | Li = 0 | % | Pr. | % | Li = 1 | Li = 0 |
|---|---|---|---|---|---|---|---|---|
| MLP 9:9-6-1:1  | Li = 1 | 29117 | 19490 | 81.8 | Sn | 81.9 | 9729 | 6426 |
| | Li = 0 | 6481 | 87304 | 81.7 | Sp | 81.9 | 2137 | 29172 |
| LNN 12:12-1:1  | Li = 1 | 22320 | 39543 | 62.70 | Sn | 63.22 | 7502 | 13302 |
| | Li = 0 | 13278 | 67251 | 63.0 | Sp | 62.63 | 4364 | 22296 |

Pr. = Parameter, Sp = Specificity, Sn =Sensitivity. Columns: Observed classifications; Rows: Predicted classifications

**Table 4.** Average values $J_k(i)_{org.avg}$ *vs.* some classic parameters of selected BINs of ecosystems

| BIN | Name | $J_0$ | $J_1$ | $J_2$ | $J_3$ | $J_4$ | $J_5$ | $Z_{in}$ | $Z_{out}$ | W |
|---|---|---|---|---|---|---|---|---|---|---|
| IWBD01 | Anemone-fish | 2.820 | 3.809 | 3.890 | 3.898 | 3.899 | 3.899 | 2.222 | 2.222 | 2.722 |
| IWBD02 | Aishihik lake | 2.752 | 3.894 | 3.988 | 4.001 | 4.004 | 4.005 | 2.167 | 2.167 | 2.667 |
| IWBD03 | Cold lake | 2.566 | 3.306 | 3.486 | 3.557 | 3.593 | 3.615 | 1.820 | 1.820 | 2.320 |
| IWBD04 | Lake of the woods | 2.803 | 4.141 | 4.320 | 4.373 | 4.397 | 4.409 | 2.194 | 2.194 | 2.694 |
| IWBD05 | Mcgregor river | 2.506 | 3.255 | 3.464 | 3.548 | 3.589 | 3.612 | 1.754 | 1.754 | 2.254 |
| IWBD06 | Parsnip river | 2.918 | 3.747 | 3.925 | 3.986 | 4.013 | 4.027 | 2.257 | 2.257 | 2.757 |
| IWBD07 | Bay lake huron | 3.028 | 4.162 | 4.274 | 4.298 | 4.308 | 4.314 | 2.431 | 2.431 | 2.931 |
| IWBD08 | Smallwood reservoir | 2.315 | 3.369 | 3.476 | 3.494 | 3.501 | 3.504 | 1.710 | 1.710 | 2.210 |
| IWBD09 | Bluthgen 2004 | 3.679 | 4.385 | 4.424 | 4.426 | 4.426 | 4.426 | 3.098 | 3.098 | 3.598 |
| IWBD10 | Davidson et al 1989 | 2.180 | 2.443 | 2.527 | 2.553 | 2.562 | 2.565 | 0.792 | 0.792 | 1.292 |
| IWBD11 | Davidson & Fisher | 2.169 | 2.087 | 2.063 | 2.056 | 2.054 | 2.054 | 1.300 | 1.300 | 1.800 |
| IWBD12 | Fonseca&ganade | 2.066 | 2.287 | 2.353 | 2.377 | 2.388 | 2.395 | 1.171 | 1.171 | 1.671 |
| IWBD13 | Joern 1979 marathon | 2.936 | 2.853 | 2.848 | 2.847 | 2.848 | 2.848 | 2.218 | 2.218 | 2.718 |
| IWBD14 | Joern 1979 altuda | 3.151 | 3.048 | 3.053 | 3.055 | 3.056 | 3.057 | 2.486 | 2.486 | 2.986 |
| IWBD15 | Leather 1991 finland | 2.031 | 5.848 | 6.313 | 6.391 | 6.411 | 6.419 | 1.377 | 1.377 | 1.877 |
| IWBD16 | Leather 1991 britain | 1.760 | 6.016 | 6.097 | 6.099 | 6.100 | 6.100 | 1.234 | 1.234 | 1.734 |
| IWBD17 | Arroyo i | 2.762 | 3.037 | 3.134 | 3.173 | 3.192 | 3.203 | 2.011 | 2.011 | 2.511 |
| IWBD18 | Arroyo ii | 2.570 | 2.976 | 3.046 | 3.063 | 3.068 | 3.070 | 1.895 | 1.895 | 2.395 |
| IWBD19 | Arroyo iii | 2.212 | 2.133 | 2.155 | 2.170 | 2.179 | 2.185 | 1.319 | 1.319 | 1.819 |
| IWBD20 | Barret & Helenurm | 2.228 | 4.124 | 4.789 | 5.108 | 5.278 | 5.374 | 1.465 | 1.465 | 1.965 |
| IWBD21 | Clements 1923 | 3.050 | 5.099 | 5.305 | 5.362 | 5.388 | 5.401 | 2.488 | 2.488 | 2.988 |
| IWBD22 | Dupont et al 2003 | 2.847 | 3.750 | 3.923 | 3.984 | 4.013 | 4.031 | 2.163 | 2.163 | 2.663 |
| IWBD23 | Elberling & Olesen | 2.352 | 3.422 | 3.693 | 3.815 | 3.879 | 3.917 | 1.688 | 1.688 | 2.188 |
| IWBD24 | Hocking 1968 | 2.193 | 3.904 | 4.109 | 4.186 | 4.226 | 4.251 | 1.600 | 1.600 | 2.100 |
| IWBD25 | Kato et al 1990 | 2.336 | 5.292 | 6.113 | 6.415 | 6.550 | 6.621 | 1.566 | 1.566 | 2.066 |

**Table 5.** Linear vs. Non-linear MIANN models of BINs of 73 Ecosystems

| MIANN models | Li | Li = 1 | Li = 0 | % | Pr. | % | Li = 1 | Li = 0 |
|---|---|---|---|---|---|---|---|---|
| MLP 12:12-9-1:1 | Li = 1 | 3992 | 2684 | 81.6 | Sn | 81.3 | 1326 | 898 |
|  | Li = 0 | 902 | 11934 | 81.6 | Sp | 81.6 | 305 | 3975 |
| LNN 11:11-1:1 | Li = 1 | 3406 | 4435 | 69.6 | Sn | 69.3 | 1131 | 1470 |
|  | Li = 0 | 1488 | 10183 | 69.7 | Sp | 69.8 | 500 | 3403 |

Pr. = Parameter, Sp = Specificity, Sn =Sensitivity. Columns: Observed classifications; Rows: Predicted classifications

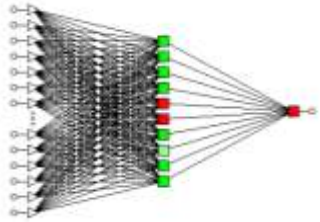
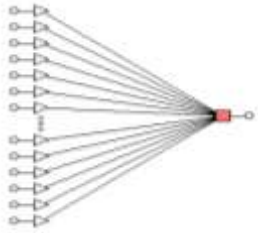
SNA can be used to how central a case is to law at the Court and measure other legal concepts. However, the model is unable to predict the future evolution of these citations. In this type of situation, application of a model able to predict the future evolution of connectivity patterns $\mathbf{L}_i$ (direct and indirect citation patterns) of different cases along time may become a useful tool. For instance, our group have reported similar models for Spanish financial law network [52]. In this sense, it is straightforward to realise that here we should use TIs to describe the complex network data. The same authors, have used different node centralities to study this network before a detect the more relevant cases (higher authority) at different times[5]. In addition, we should use a time-series technique if we want to predict the future evolution of case citation patterns. Last, as the problem is very probably no-linear we should consider the probability of use a powerful non-linear algorithm to fit the data, as is the case of ANNs. All these features are present in MIANN models. Consequently, we decided to combine the new $J_k(i)$ centralities with MIANN analysis to model this data. First, we calculated the $J_k(i)$ values of all cases. The **Table 6** summarises some of these values in two different time scales. In this table we give the average of $J_k(i)$ values for USSC in 5Ks cites slots *vs.* decades of the citing case. Here 5K cites slot means that we construct a new sub-network for each period of time from the starting data in which appear 5000 more cites to U.S. Supreme Court cases. It means that 5K scale is divided in irregular periods of time but all sub-networks has the same number of links m = 5000 and different number of nodes (cases). Conversely, natural time scale networks contain different numbers of nodes (cases) and links (cites) accumulated in regular periods of time (decades).

Previously to carry out the MIANN analysis we decided to explore the possible relationships between the $J_k(i)$ values and other node centralities previously used to describe this network. We carried out a TWJCA of this dataset as well; see **Figure 2 (C)**. We used a Number of variables = 15 and Number of cases = 25 decades. The other variables used, in addition to $J_k(i)$ values, were the following: $Z_{in}$ = in degree, $Z_{out}$ = out degree, Hub = raw Hub score, Hubrank = rank of hub score, Auth = raw Authority score, Authrank = Rank of Authority score, Betweeness = 5KCN Node betweeness centrality for a give case, Eigenvector = eigenvector centrality measure.

**Table 6.** Average of USSC $J_k(i)$ values in two different time scales: 5Ks *vs.* decades of the citing case

| Decade | $J_0$ | $J_1$ | $J_2$ | $J_3$ | $J_4$ | $J_5$ | 5Ks | $J_0$ | $J_1$ | $J_2$ | $J_3$ | $J_4$ | $J_5$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1780 | 8.3 | 10.5 | 11.1 | 11.3 | 11.3 | 11.4 | (000K-005K) | 19.2 | 17.4 | 14.9 | 13.8 | 13.3 | 13.0 |
| 1790 | 10.7 | 13.8 | 14.4 | 14.7 | 14.7 | 14.8 | (005K-010K) | 13.0 | 10.4 | 9.3 | 8.9 | 8.8 | 8.7 |
| 1800 | 12.7 | 16.1 | 16.8 | 17.1 | 17.2 | 17.2 | (010K-015K) | 8.7 | 6.7 | 6.1 | 5.9 | 5.8 | 5.8 |
| 1810 | 14.1 | 17.6 | 18.2 | 18.3 | 18.4 | 18.4 | (015K-020K) | 10.6 | 8.2 | 7.3 | 7.0 | 7.0 | 6.9 |
| 1820 | 12.4 | 15.6 | 16.0 | 16.1 | 16.1 | 16.1 | (020K-025K) | 23.5 | 17.7 | 15.5 | 14.9 | 14.7 | 14.7 |
| 1830 | 9.3 | 11.3 | 11.4 | 11.3 | 11.3 | 11.3 | (025K-030K) | 15.9 | 12.6 | 11.1 | 10.7 | 10.5 | 10.5 |
| 1840 | 8.6 | 10.5 | 10.4 | 10.3 | 10.3 | 10.2 | (030K-035K) | 25.1 | 19.4 | 17.0 | 16.3 | 16.0 | 15.8 |
| 1850 | 7.7 | 8.8 | 8.4 | 8.2 | 8.1 | 8.0 | (035K-040K) | 10.6 | 8.4 | 7.5 | 7.3 | 7.1 | 7.1 |
| 1860 | 9.1 | 10.7 | 10.3 | 10.0 | 9.9 | 9.9 | (040K-045K) | 25.1 | 19.4 | 17.0 | 16.3 | 16.0 | 15.8 |
| 1870 | 9.3 | 9.7 | 9.3 | 9.1 | 9.1 | 9.0 | (045K-050K) | 10.6 | 8.4 | 7.5 | 7.3 | 7.1 | 7.1 |
| 1880 | 10.9 | 9.9 | 9.3 | 9.1 | 9.1 | 9.0 | (050K-055K) | 22.4 | 16.2 | 14.4 | 13.9 | 13.8 | 13.7 |
| 1890 | 14.0 | 12.0 | 11.0 | 10.7 | 10.6 | 10.6 | (055K-060K) | 8.2 | 6.2 | 5.7 | 5.5 | 5.5 | 5.4 |
| 1900 | 13.5 | 12.6 | 12.1 | 11.9 | 11.9 | 11.9 | (060K-065K) | 18.4 | 13.4 | 12.1 | 11.7 | 11.6 | 11.5 |
| 1910 | 17.4 | 15.0 | 13.6 | 13.2 | 13.1 | 13.0 | (065K-070K) | 30.5 | 21.4 | 19.0 | 18.2 | 18.0 | 17.9 |
| 1920 | 18.0 | 14.3 | 13.1 | 12.8 | 12.7 | 12.6 | (070K-075K) | 12.5 | 8.9 | 7.9 | 7.7 | 7.6 | 7.6 |
| 1930 | 25.0 | 16.9 | 14.9 | 14.4 | 14.2 | 14.1 | (075K-080K) | 25.3 | 17.4 | 15.8 | 15.3 | 15.2 | 15.1 |
| 1770 | 2.2 | 2.4 | 2.4 | 2.4 | 2.4 | 2.4 | (080K-085K) | 38.7 | 26.5 | 23.8 | 23.3 | 23.2 | 23.1 |
| 1940 | 26.7 | 16.7 | 14.3 | 13.8 | 13.6 | 13.5 | (085K-090K) | 21.9 | 16.3 | 14.4 | 13.7 | 13.4 | 13.3 |
| 1950 | 19.9 | 14.3 | 12.9 | 12.6 | 12.4 | 12.3 | (090K-095K) | 25.3 | 17.4 | 15.9 | 15.5 | 15.3 | 15.2 |
| 1960 | 25.6 | 18.4 | 15.4 | 14.5 | 14.2 | 14.1 | (095K-100K) | 30.2 | 20.8 | 18.4 | 17.9 | 17.8 | 17.7 |
| 1760 | 43.7 | 45.1 | 45.4 | 45.5 | 45.5 | 45.5 | (100K-105K) | 26.7 | 18.6 | 16.8 | 16.3 | 16.1 | 16.0 |
| 1970 | 22.3 | 12.4 | 10.3 | 9.8 | 9.6 | 9.5 | (105K-110K) | 12.6 | 9.3 | 8.4 | 8.2 | 8.2 | 8.1 |
| 1980 | 17.3 | 8.1 | 6.7 | 6.4 | 6.3 | 6.3 | (110K-115K) | 21.2 | 15.1 | 13.7 | 13.3 | 13.2 | 13.1 |
| 1990 | 27.1 | 8.0 | 4.7 | 4.1 | 3.9 | 3.9 | (115K-120K) | 22.8 | 16.6 | 14.8 | 14.3 | 14.0 | 13.9 |
| 2000 | 64.4 | 9.2 | 2.6 | 1.0 | 0.5 | 0.3 | (120K-125K) | 26.0 | 19.6 | 17.5 | 16.9 | 16.7 | 16.6 |

**Table 7.** Linear vs. Non-linear MIANN models of 43 5KCNs of the U.S. Supreme Court

| MIANN models | Li | Li = 1 | Li = 0 | % | Pr. | % | Li = 1 | Li = 0 |
|---|---|---|---|---|---|---|---|---|
| MLP 18:18-10-1:1  | Li = 1 | 81225 | 51008 | 82.49 | Sn | 82.76 | 26985 | 17014 |
| | Li = 0 | 16917 | 243415 | 82.66 | Sp | 82.7 | 5728 | 81128 |
| LNN 18:18-1:1  | Li = 1 | 77871 | 60826 | 79.33 | Sn | 79.35 | 25950 | 20284 |
| | Li = 0 | 20271 | 233597 | 79.33 | Sp | 79.3 | 6763 | 77858 |

Pr. = Parameter, Sp = Specificity, Sn =Sensitivity. Columns: Observed classifications; Rows: Predicted classifications

We found a Number of blocks = 112 with a Threshold computed from data = 0.4898979 (StDv/2), Mean = -0.00, and Standard Deviation = 0.9797958. TWJCA shows that $J_k(j)$ indices of the U.S. Supreme Court with degree k = 1 to 5 form their own cluster. The effect is stronger in the decades from 1840 to 1990. It is interesting that $J_0(j)$ does not forms clusters with the other $J_k(j)$ indices. In general $J_k(j)$ indices do not form clusters with other TIs or node centralities of the U.S. Supreme Court. The MIANN model here was also the MLP, now with slightly better goodness-of-fit with respect (Ac, Sn, and Sp approximately 82-83%) to MRNs and BINs (Ac, Sn, and Sp approximately 81%). Interestingly, the best linear model is notably better here with (Ac, Sn, and Sp approximately 79%), which is lower but similar to the MLP. Considering that both models MLP 18:18-10-1:1 and LNN 18:18-1:1 has very similar performance and the same number of inputs we should consider the simpler LNN (18 variables and 0 hidden neurons) model also as a very good model. The MLP needs 10 hidden neurons to increase performance only in 1-2%. The situation is curious, linear models were increasing in performance from bio-molecular process to ecological and social systems.

## 6. CONCLUSIONS

In this work, we report by the frist time a new class of Balaban type parameters called Balaban-Markov centralities. Contrary to tradition, the new indices were not defined for small molecules only but for all classes of systems susceptible to be represented for graphs. We also report three proof-of-concept experiments, to test the power of the new $J_k(i)$ indices to predict actual node connectivity patterns in complex bio-molecular, ecological, and social networks.

## 7. ACKNOWLEDGMENTS

The authors acknowledge the kind attention of the Editor-In-Chief Prof. Alessandro Giuliani Istituto Superiore di Sanità (Italian NIH) Environment and Health Dept Rome, Italy.

## REFERENCES

1. Boccaletti S, Latora V, Moreno Y et al. Complex networks: Structure and dynamics, Physics Reports 2006;424:175-308.
2. Jeong H, Tombor B, Albert R et al. The large-scale organization of metabolic networks, Nature 2000;407:651-654.
3. Estrada E. Food webs robustness to biodiversity loss: the roles of connectance, expansibility and degree distribution., J Theor Biol 2007;244:296-307.
4. Fowler JH, Jeon S. The authority of Supreme Court precedent, Social Networks 2008; 30:16-30.
5. Fowler JH, Johnson TR, II JFS et al. Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court 2007.
6. Estrada E. Returnability as a criterion of disequilibrium in atmospheric reactions network, Journal of Mathematical Chemistry 2012;50:1363-1372.
7. Rosenquist JN, Murabito J, Fowler JH et al. The spread of alcohol consumption behavior in a large social network, Ann Intern Med 2010;152:426-433, w141.
8. Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years, N Engl J Med 2007;357:370-379.
9. Bornholdt S, Schuster HG. Handbook of Graphs and Complex Networks: From the Genome to the Internet. Wheinheim: WILEY-VCH GmbH & CO. KGa., 2003.
10. Sporns O, Tononi G. Classes of network connectivity and dynamics, Complexity;7:28-38.
11. Estrada E, Uriarte E. Recent advances on the role of topological indices in drug discovery research., Curr Med Chem 2001;8:1573-1588.
12. Garcia-Domenech R, Galvez J, de Julian-Ortiz JV et al. Some new trends in chemical graph theory, Chem Rev 2008;108:1127-1169.
13. Komosinski M, Kubiak M. Quantitative measure of structural and geometric similarity of 3D morphologies, Complexity;16:40-52.

14. Diudea MV, Aleksander I, Kurt et al. Network analysis using a novel highly discriminating topological index, Complexity;16:32-39.
15. Dehmer M, Department for Biomedical Informatics and Mechatronics IfBaTR, UMIT, Eduard Wallnoefer Zentrum 1, A-6060, Hall in Tyrol, Austria, Department for Biomedical Informatics and Mechatronics IfBaTR, UMIT, Eduard Wallnoefer Zentrum 1, A-6060, Hall in Tyrol, Austria et al. Generalized graph entropies, Complexity;17:45-50.
16. T BA. HIGHLY DISCRIMINATING DISTANCE-BASED TOPOLOGICAL INDEX, Chem. Phys. Lett. 1981;89:399-404.
17. Basak SC, Balaban AT, Grunwald GD et al. Topological indices: their nature and mutual relatedness, J Chem Inf Comput Sci 2000;40:891-898.
18. Wiener H. Correlation of heats of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons, Journal of the American Chemical Society 1947;69:2636–2638.
19. Wiener H. Relation of the physical properties of the isomeric alkanes to molecular structure. Surface tension, specific dispersion, and critical solution temperature in aniline, The Journal of Physical and Colloid Chemistry 1948;52:1082-1089.
20. Wiener H. Vapor pressure–temperature relationships among the branched paraffin hydrocarbons, The Journal of Physical and Colloid Chemistry 1948;52:425-430.
21. Hosoya H. Topological index,a newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons, Bulletin of the Chemical Society of Japan 1971;44:2332-2339.
22. Hosoya H. Mathematical meaning and importance of the topological index Z, Croatica Chemica Acta 2007;80:239-249.
23. Randic M. Characterization of molecular branching, J. Am. Chem. Soc. 1975;97:6609–6615.
24. Kier LB, Hall LH, Murray WJ et al. Molecular connectivity. I: Relationship to nonspecific local anesthesia, J Pharm Sci 1975;64:1971-1974.
25. Hall LH, Kier LB. Issues in representation of molecular structure the development of molecular, J Mol Graph Model 2001;20:4-18.
26. Balaban AT. From chemical graphs to 3D molecular modeling. In: Balaban A. T. (ed) From chemical topology to three-dimensional geometry. New York: Plenum Publishers, 1997, 420.
27. Sharma SK, Kumar P, Narasimhan B et al. Synthesis, antimicrobial, anticancer evaluation and QSAR studies of, Eur J Med Chem 2012;48:16-25.
28. Thakur A, Thakur M, Khadikar PV et al. QSAR study on benzenesulphonamide carbonic anhydrase inhibitors: topological, Bioorg Med Chem 2004;12:789-793.
29. Krawczuk A, Voelkel A, Lulek J et al. Use of topological indices of polychlorinated biphenyls in structure-retention, J Chromatogr A 2003;1018:63-71.
30. Yadav S, Kumar P, De Clercq E et al. 4-[1-(Substituted aryl/alkyl carbonyl)-benzoimidazol-2-yl]-benzenesulfonic acids, Eur J Med Chem 2010;45:5985-5997.
31. Naik PK, Dubey A, Kumar R. Development of predictive quantitative structure-activity relationship models of, J Biomol Screen 2010;15:1194-1203.
32. Fernandes JP, Pasqualoto KF, Felli VM et al. QSAR modeling of a set of pyrazinoate esters as antituberculosis prodrugs, Arch Pharm (Weinheim) 2010;343:91-97.
33. Panaye A, Doucet JP, Devillers J et al. Decision trees versus support vector machine for classification of androgen, SAR QSAR Environ Res 2008;19:129-151.
34. Zavrsnik D, Muratovic S, Spirtovic S. QSAR and QSPR study of derivatives 4-arylaminocoumarin, Bosn J Basic Med Sci 2003;3:59-63.
35. Ma XL, Chen C, Yang J. Predictive model of blood-brain barrier penetration of organic compounds, Acta Pharmacol Sin 2005;26:500-512.
36. Dashtbozorgi Z, Golmohammadi H. Quantitative structure-property relationship modeling of water-to-wet butyl, J Sep Sci 2010;33:3800-3810.

37. Jalali-Heravi M, Fatemi MH. Prediction of thermal conductivity detection response factors using an artificial neural network, J Chromatogr A 2000;897:227-235.
38. Dehmer M, Grabner M, Varmuza K. Information indices with high discriminative power for graphs, PLoS One 2012;7:e31214.
39. Basak SC, Mills D, Gute BD et al. Use of mathematical structural invariants in analyzing combinatorial libraries: a, Curr Comput Aided Drug Des 2010;6:240-251.
40. Ratkiewicz A, Truong TN. Application of chemical graph theory for automated mechanism generation, J Chem Inf Comput Sci 2003;43:36-44.
41. Rastija V, Medic-Saric M. QSAR study of antioxidant activity of wine polyphenols, Eur J Med Chem 2009;44:400-408.
42. Randic M, Pompe M. The variable molecular descriptors based on distance related matrices, J Chem Inf Comput Sci 2001;41:575-581.
43. Mueller LA, Kugler KG, Graber A et al. Structural Measures for Network Biology Using QuACN, BMC Bioinformatics 2011;12:492.
44. Junker BH, Koschutzki D, Schreiber F. Exploration of biological network centralities with CentiBiN, BMC Bioinformatics 2006;7:219.
45. Batagelj V, Mrvar A. Pajek— Analysis and Visualization of Large Networks, Lecture Notes in Computer Science 2002;2265:477-478.
46. Mueller LA, Kugler KG, Dander A et al. QuACN: an R package for analyzing complex biological networks quantitatively, Bioinformatics 2011;27:140-141.
47. Balaban AT, Khadikar PV, Supuran CT et al. Study on supramolecular complexing ability vis-a-vis estimation of pKa of, Bioorg Med Chem Lett 2005;15:3966-3973.
48. Gonzalez-Diaz H, Prado-Prado F, Ubeira FM. Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach, Curr Top Med Chem 2008;8:1676-1690.
49. Gonzalez-Diaz H, Riera-Fernandez P, Pazos A et al. The Rucker-Markov invariants of complex Bio-Systems: applications in Parasitology and Neuroinformatics, Biosystems 2013;111:199-207.
50. Riera-Fernandez P, Munteanu CR, Martin-Romalde R et al. Markov-Randic Indices for QSPR Re-Evaluation of Metabolic, Parasite-Host, Fasciolosis Spreading, Brain Cortex and Legal-Social Complex Networks, Current Bioinformatics 2013;8:401-415.
51. Gonzalez-Diaz H, Riera-Fernandez P. New Markov-autocorrelation indices for re-evaluation of links in chemical and biological complex networks used in metabolomics, parasitology, neurosciences, and epidemiology, J Chem Inf Model 2012;52:3331-3340.
52. Riera-Fernandez P, Munteanu CR, Escobar M et al. New Markov-Shannon Entropy models to assess connectivity quality in complex networks: from molecular to cellular pathway, Parasite-Host, Neural, Industry, and Legal-Social networks, J Theor Biol 2012;293:174-188.
53. Hill T, Lewicki P. STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining. Tulsa: StatSoft, 2006.
54. Frank E, Hall M, Trigg L et al. Data mining in bioinformatics using Weka. Bioinformatics. England, 2004, 2479-2481.
55. Gonzalez-Diaz H, Arrasate S, Sotomayor N et al. MIANN Models in Medicinal, Physical and Organic Chemistry, Curr Top Med Chem 2013;13:619-641.
56. Box GEP, Jenkins GM. Time series analysis. Holden-Day, 1970.
57. Hill T, Lewicki P. STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining. Tulsa: StatSoft, 2006
58. Riera-Fernández P, Munteanu CR, Pedreira-Souto N et al. Definition of Markov-Harary Invariants and Review of Classic Topological Indices and Databases in Biology, Parasitology, Technology, and Social-Legal Networks, Current Bioinformatics 2011;6:94-121.
59. Riera-Fernandez P, Munteanu CR, Pedreira-Souto N et al. Definition of Markov-Harary Invariants and Review of Classic Topological Indices and Databases in Biology, Parasitology, Technology, and Social-Legal Networks, Current Bioinformatics 2011;6:94-121.

60. Estrada E. Information mobility in complex networks., Phys Rev E Stat Nonlin Soft Matter Phys 2009;80:026104.
61. Mazurie A, Bonchev D, Schwikowski B et al. Evolution of metabolic network organization, BMC Syst Biol 2010;4:59.
62. Kier LB, Bonchev D, Buck GA. Modeling biochemical networks: a cellular-automata approach, Chem Biodivers 2005;2:233-243.
63. Lee DS, Park J, Kay KA et al. The implications of human metabolic network topology for disease comorbidity, Proceedings of the National Academy of Sciences of the United States of America 2008;105:9880-9885.
64. Wasserman S, Faust K. Social network analysis: methods and applications. Cambridge: Cambridge University Press, 1999.

**Artículo 4.**

**Predicción de Redes Bio-Moleculares, Socio-Económicas y Jurídicas con MI-NODES**

MI-NODES Multiscale Models of Metabolic Reactions, Ecological, Brain Connecto me, Epidemic, World Trade, and Legal-Social Ne tworks. A. Duardo-Sanchez, H. González-Díaz, A. Pazos *Current Bioinformatics*, **2014**, 9, *en imprenta*.

*Send Orders of Reprints at reprints@benthamscience.net*

***Current Bioinformatics*, 2014, *9*, 000 - 000**

# MI-NODES Multiscale Models of Metabolic Reactions, Brain Connectome, Ecological, Epidemic, World Trade, and Legal-Social Networks

Aliuska Duardo-Sanchez [1,2], Humberto González-Díaz [3,4,*] and Alejandro Pazos[1]

[1] *Department of Information and Communication Technologies, University of A Coruña UDC, 15071, A Coruña, Spain.*

[2] *Department of Especial Public Law, Financial and Tributary Law Area, Faculty of Law, University of Santiago de Compostela, 15782, Santiago de Compostela, Spain*

[3] *Department of Organic Chemistry II, University of the Basque Country UPV/EHU, 48940, Leioa, Spain.*

[4] *IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain.*

**Abstract:** Complex systems and networks appear in almost all areas of reality. We find then from proteins residue networks to Protein Interaction Networks (PINs). Chemical reactions form Metabolic Reactions Networks (MRNs) in living beings or Atmospheric reaction networks in planets and moons. Network of neurons appear in the worm *C. elegans*, in Human brain connectome, or in Artificial Neural Networks (ANNs). Infection spreading networks exist for contagious outbreaks networks in humans and in malware epidemiology for infection with viral software in internet or wireless networks. Social-legal networks with different rule evolved from swarm intelligence, to hunter-gathered societies, or citation networks of U.S. Supreme Court. In all these cases, we can see the same question. Can we predict the links based on structural information? We propose to solve the problem using Quantitative Structure-Property Relationship (QSPR) techniques commonly used in chemo-informatics. In so doing, we need software able to transform all types of networks/graphs like drug structure, drug-target interactions, protein structure, protein interactions, metabolic reactions, brain connectome, or social networks into numerical parameters. Consequently, we need to process in alignment-free mode multitarget, multiscale, and multiplexing, information. Later we have to seek the QSPR model with Machine Learning techniques. MI-NODES / MARCH-INSIDE: Markov Chain Invariants for Complex Networks Simulation and Design is this type of software. Here we review the evolution of the software from chemo-informatics to bioinformatics and systems biology. This is an effort to develop a universal tool to study structure-property relationships in complex systems.

***Keywords*:** QSPR models in Complex Networks; Drug-Target Networks; Metabolic networks; Brain Connectome; Social networks; World Trade; US Supreme Court citation networks; Spain's Financial Law.

## 1. INTRODUCTION

### Structure-property problem in complex systems

Complex systems and networks appear in phenomena belonging to almost all areas of reality at very different temporal and spatial scales [1].

*Address correspondence to this author at: *Department of Organic Chemistry II, University of the Basque Country UPV/EHU, 48940, Leioa, Spain*; E-mail: Humberto.gonzalezdiaz@ehu.es

We find them from bio-molecular structures of proteins residue networks [2] to Protein Interaction Networks (PINs)[3]. The coupling of chemical reactions lead to the formation of Metabolic Reactions Networks (MRNs) [4, 5] in living beings or Atmospheric Reaction Networks (ARNs) in planets and moons like Earth, Mars, Venus, and Titan [6]. Complex patterns appear in the network of neurons of the worm *C. elegans* [7], in Human

1574-8936/14 $58.00+.00 

Brain Connectome [8], or in one Artificial Neural Networks (ANNs) [9]. Spreading patterns appears for contagious outbreaks networks in humans [10] or in malware epidemiology due to infection with viral software in internet or Wi-Fi wireless networks [11, 12]. Complex behavior emerge from basic rules of Swarm Intelligence (SI) [13], collaboration in hunter-gathered societies [14], or in legislation code in the citation network of U.S. Supreme Court (USSC), as well [15]. The most basic issues are structural: how does one characterize the connectivity patterns in those networks? Are there any unifying features underlying their topology? Different research groups have begun to shed light over these unifying aspects of the structure and dynamics of complex networks indeed [1, 7, 16-21]. Networks are represented by means of a graph as a way to capture essential information. Graphs in turn are sets of items, drawn as dots, or *nodes*, interconnected by lines or arcs, which represents wires, ties, links, edges, bonds, or in general pair-wise relationships. Consequently, the nodes can represent atoms, molecules, proteins, nucleic acids, drugs, cells, organisms, parasites, people, words, laws, computers or any other part of a real system. Moreover, the links represent relationships between the nodes such as chemical bonds, physical interactions, metabolic pathways, pharmacological action, law recurrence or social ties [4, 22-30].

Tenazinha and Vinga [31] reviewed frameworks currently available for modeling and analyzing integrated biological networks, in particular metabolic, gene regulatory and signaling networks. In effect, there are different experimental and/or theoretical methods to assign node-node links depending on the type of network we want to construct. Unfortunately, many of these methods are expensive in terms of time or resources (especially the experimental ones). In addition, different methods to link nodes in the same type of network are not very accurate in such a way that they do not always coincide. One possible solution to this problem is the use of Quantitative Structure-Property (QSPR) models. This methodology has been traditionally used in chemo-informatics. Most often, QSPR-like models use as input structural parameters derived from the graph representation of the network-like system under study [32]. Many authors refer to the numerical parameters of a graph as Topological Indices (TIs); mainly in the case of global studies (properties of full system). We can use local TIs of a sub-graph or centralities $C_t(j)$ of type t for the $j^{th}$ node in the network; if we want to study a local property of a given part of the system [33-35]. In **Table 1**, we give the names, symbol, formula, and software used to calculate some of these centralities.

In order to develop such computational models we need to use modeling techniques to process chemical information from public databases. These databases have accumulated immense datasets of experimental results of pharmacological trials for many compounds. For instance, STITCH [36-38], TTD [39], SuperTarget [40, 41], or the colossal CHEMBL lists thousands of drugs, targets, and drug-target interactions. This huge amount of information offers a fertile field for the application of computational techniques [42, 43]. The analysis of all this data is very complex due to different features of the chemical and pharmacological information present: (1) multi-scaling, (2) multi-targeting, (3) alignment dependent, and/or (4) multi-output or plexing nature. The same features appear in biological, technological, social, and other complex networks.

**Why do we need multiscale models?**

One of the more important characteristics enumerated before is the multi-scale nature of many important problems. Currently, the use of QSPR-like models in which the inputs are graph parameters is not limited to the study of molecules and has been extended to other complex systems [44]. As we mentioned in the previous paragraph, in multi-target modelling we need to incorporate information about the drug and different molecular targets (proteins, RNA, gene). In this case, we can solve the problem using molecular descriptors. However, in the case of not molecular complex networks we are out of the chemical scale. We can find complex systems formed by networks in many different scales. In general, these scales may be classified as time and spatial scales. In the case of time scales, we can find different dynamic networks in a same or different problem that change the pattern of links in different time scales (seconds, min, hours, days, years, or seasons). In this case, we can still circumvent the problem with MA models like those of Bob and Jenkins mentioned before [45].

**Table 1.** MI-NODES *vs.* some classic node centralities

| Name | Formula | Software | Ref. |
|---|---|---|---|
| Degree | $C_{\deg}(j)=\deg(j)$ | CBI | [34] |
| Eccentricity | $C_{ecc}(j)=\max\{dist(i,j)\}^{-1}$ | CBI | |
| Closeness | $C_{clo}(j)=\left(\sum_{j\in V} dist(i,j)\right)^{-1}$ | CBI | |
| Radiality | $C_{rad}(j)=\sum_{w\in V}(\Delta_G+1-dist(i,j))/(n-1)$ | CBI | |
| Centroid Values | $C_{cen}(j)=\min\{f(i,j): i\in V\setminus\{j\}\}$ | CBI | |
| Stress | $C_{str}(j)=\sum_{s\neq v\in V}\sum_{t\neq v\in V}\sigma_{st}(j)$ | CBI | |
| Shortest-path Betweenness | $C_{spb}(j)=\sum_{s\neq j\in V}\sum_{t\neq j\in V}\delta_{st}(j)$ | CBI | |
| Current-Flow Closeness | $C_{cfc}(j)=(n-1)/\left(\sum_{i\neq V} p_{ji}(j)-p_{ij}(i)\right)$ | CBI | |
| Current-Flow Betweenness | $C_{cfb}(j)=\sum_{s,t\in V}\tau_{st}(j)/(n-1)(n-2)$ | CBI | |
| Katz Status Index | $C_{katz}=\sum_{k=1}^{\infty}\alpha^k\cdot(\mathbf{A}^t)^k\cdot\mathbf{u}$ | CBI | |
| Eigenvector | $EC(j)=e_1(j)$ | CBI | |
| Closeness Vitality | $C_{clv}(j)=W(G)-W(G\setminus\{j\})$ | CBI | |
| Markov-Randic | ${}^kC_{\chi}(j)=\sum_{i}^{\delta_j}(\delta_i\cdot\delta_j)^{1/2}\cdot{}^kp_{ij}$ | MI | [46] |
| Markov-Shannon entropy | ${}^kC_{\theta}(j)=-\sum_{i}^{n}({}^kp_j)\cdot\log({}^kp_j)$ | MI | [47] |
| Markov Spectral moments | ${}^kC_{\pi}(j)=\sum_{i=j}^{n}{}^kp_{ij}=Tr\left[({}^1\mathbf{\Pi})^k\right]$ | MI | [48] |
| Markov-Harary | ${}^kC_H(j)=\frac{1}{2}\sum_{i}^{\delta_j}{}^kp_{ij}^{-1}$ | MI | [49] |
| Markov-Galvez | ${}^kC_G(j)=\frac{1}{2}\sum_{i,j}^{n}\left|{}^kCT_{ij}\right|\cdot\delta_j$ | MI | [50] |
| Markov-Rucker | ${}^kC_{WC}(j)=\frac{1}{2}\sum_{i}^{\delta_j}{}^kp_{ij}$ | MI | [51] |
| Markov-BM Autocorrelation | ${}^kC_{BM}(j)=\frac{1}{2}\sum_{i}^{\delta_j}{}^kp_{ij}\cdot{}^kp_{ji}$ | MI | [52] |
| Markov-Wiener | ${}^kC_w(j)=\frac{1}{2}\cdot\sum_{i\to 1}^{\delta_j}{}^kp_{ij}\cdot\mathrm{d}_{ij}$ | MI | [53] |
| Markov-Balaban | ${}^kC_J(j)=\frac{q}{\mu+1}\cdot\sum_{i\to j}^{\delta j}\left({}^kp_{ij}\cdot S_i\cdot S_j\right)^{-1/2}$ | MI | - |

[a] All symbols used in these formulae are very common in networks literature and cannot be explained in detail here. However, $G$ is an undirected or directed graph with $n = |V|$ vertices; deg($v$) denotes the degree of the vertex $v$ in an undirected graph; dist($v$, $w$) denotes the length of a shortest path between the vertices $v$ and $w$; $\sigma_{st}$ denotes the number of shortest paths from $s$ to $t$ and $\sigma_{st}(v)$ the number of shortest path from $s$ to $t$ that use the vertex $v$. **D** and **A** are the topological distance and the adjacency matrix of the graph $G$. Please, for more details see the references cited and others.

**Why do we need alignment-free models?**

Alignment-based and alignment-free methods are two fundamentally different methods used to compare sequences, and genomes by extension [54]. This approach is very useful but only when we found a high homology between the query and the template sequences deposited in the data base and therefore may fail in case of low homology [55]. The lack of function annotation (defined biological function) for the best alignment matches is another cause for alignment pitfalls [56]. Yet, functional information - either experimentally validated or computationally inferred by similarity - remains completely missing for approximately 30% of human proteins [57]. In 2012, Wood et al. [58] analyzed 1,474 prokaryotic genome annotations in GenBank. They identified 13,602 likely missed genes that are homologues to non-hypothetical proteins. It is very relevant that they also identified 11,792 likely missed genes that are homologues only to hypothetical proteins, despite evidence of their protein-coding nature. Alignment approaches also views proteins and nucleic acids as linear sequences of discrete units similar to linguistic representations ignoring 3D structure and overlooks well-documented long-range interactions [59]. On the other hand, alignment-free methods have emerged as a solution to these problems. Vinga and Almeida [60] reviewed two of the more important types of alignment free methods: (1) methods based on word frequency and resolution-free methods. In parallel, Chou [61, 62], Randic [63], González-Díaz [29, 64-70], and others have introduced alignment-free parameters for the pseudo-folding of sequences into geometrically constrained 2D, 3D, or higher dimension spaces using simple heuristics. Pseudo-folding parameters or sequence molecular descriptors codify non-linear relationships without necessity of determination of real 3D structures (graph representations) and are used as inputs of machine learning experiments to seek QSPR models able to predict function from sequence without rely upon alignment [35, 66, 71-74].

**Why do we need multitarget models?**

Multi-targeting complication emerges due to the existence of multi-target compounds [75-77], which led to the formation of complex networks of drug-target and/or target-target interactions. We can represent target interactions as networks of nodes (proteins, gene, RNAs, miRNAs) interconnected by a link when there is a target-target interaction between two of them. In addition, we can represent drug-target networks as a graph with two type of nodes drugs ($d_i$) and targets ($t_j$) interconnected by links ($L_{ij}$). Barabasi *et al.* [78], constructed a drug-target network based on Food and Drug Administration (FDA) drugs and proteins linked by drug-target binary associations. Yamanishi *et al.* [79] also reported a predictive algorithm to construct drug-target networks. Csermely *et al.* [80] have reviewed the state-of-art and trends on the use of networks, including drug-target networks, for drug discovery. In general, many of the classic models used in chemo-informatics are able to predict the biological activity of some types of drugs against only one target using molecular descriptors of the drug. An alternative is the development of general multi-target models able to predict the interaction ($L_{ij}$) of large libraries of drugs ($d_i$) with a large number of targets (multiple-target models). In this case, we can use molecular descriptors of the drug and the target. For instance, Viña *et al*. [29] and Prado-Prado *et al*. [81, 82] predicted the different drug-target network using the software MI to calculate the structural indices.

**Why do we need multiplexing models?**

However, in multi-plexing modelling we need to use additional operators to incorporate non-structural information. The non-structural information here refers to different assay conditions ($c_j$) like time, concentrations, temperature, cellular targets, tissues, organisms, *etc.* In recent works González-Díaz *et al.*, adapted the idea of Moving Average (MA) operators used in time series analysis with a similar purpose. MA models become popular after the initial works of Bob and Jenkins [45]. In multi-output modeling, we calculate the MA operators as the average of the property of the system (molecular descriptors or others) for all drugs or targets with a specific response in one assay carry out at under a sub-set of conditions ($c_j$). Consequently, our MA operator is not acting over a time domain but over a sub-set of conditions of the pharmacological assays. Botella-Rocamora *et al.* [83], have applied MA of time series theory to the spatial domain, making use of a spatial MA to define dependence on the risk of a disease

occurring. The main objective of our work is assessing links in different complex networks. For it, we use MA of properties of nodes of networks

## 2. FROM MARCH-INSIDE TO MI-NODES

In a effort to solve the previous problem, González-Díaz *et al.* introduced the software called MARCH-INSIDE (Markovian Chemicals In Silico Design), or shortly MI, which has become a very useful tool for QSPR studies for drugs, proteins, and complex systems in general [65, 84-97]. MI calculates descriptors ${}^{k}D_t(G_m)$ of type t (entropies, moments, means) and order k for all or some nodes (atoms, aminoacids, nucleic bases) using molecular graph $G_m$ of $m^{th}$ molecule. The graph G represent the 1D (sequence), 2D (secondary), or 3D (spatial) structure of a molecular system drug, protein, RNA, artificial polymers, *etc*. [9, 98]. In **Figure 1,** we illustrate the user-software interface for classic MI (top) or MI-NODES (bottom).

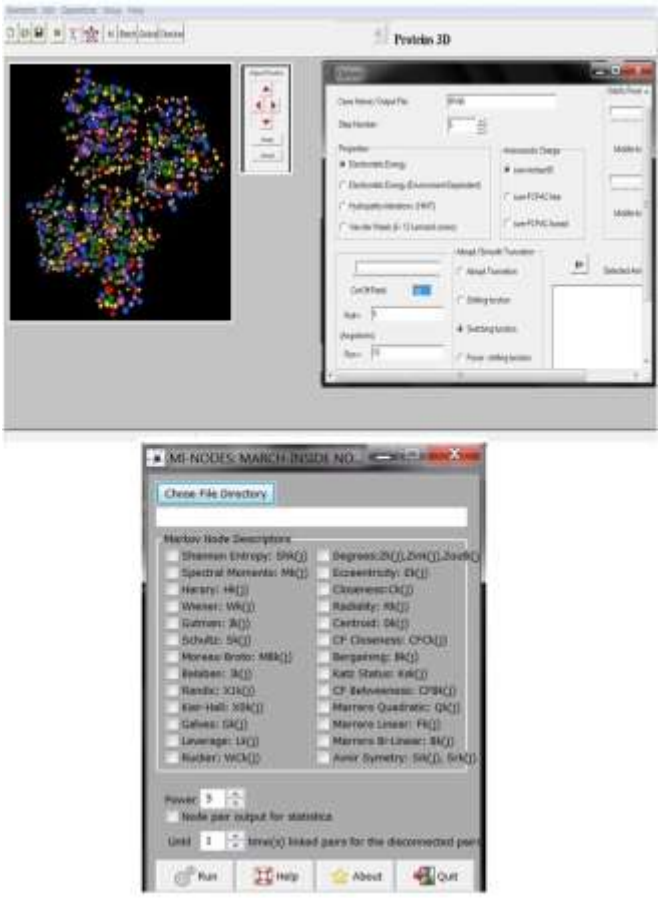



**Figure 1.** MI and MI-NODES user interfaces.

However, MI can perform a limited manage of other complex networks. Recently, we have re-programmed the MI application creating a new software application able to manage complex networks. The new program is called MI-NODES (MARCH-INSIDE for NOde DEScriptors) is able to upload files with .mat, .net, and .dat formats and is compatible with other software like Pajek [99] or CentiBin [34]. A very interesting feature of MI-NODES is that it can calculate general versions of classic molecular TIs for large complex networks using Markov Chains theory**.** In **Figure 2,** we show thee general steps used to develop a QSPR model based on the MI algorithm. Briefly, the steps of the MI algorithm are the following.

(drugs, proteins, reactions, laws, neurons, *etc.*) that form links ($L_{ij}$) in specific sub-set of conditions ($c_j$).

Step 1 - MI algorithm reads the input files with structural information of the system; essentially nodes, links, and weights.

Step 2 - MI creates a node-node connectivity or adjacency matrix **A**, if not uploaded in the input file. The elements of **A** are $a_{ij} = 1$ if the node $a_i$ is connected to the node $a_j$ and $a_{ij} = 0$ otherwise.

Step 3 - MI transforms **A** into a weighted matrix **W**. The elements of **W** are $w_{ij} = w_j$ if $a_{ij} = 1$ and $w_{ij} = 0$ otherwise. For molecules, the weights are the atomic electronegativity ($\chi_j$), polarizability ($\alpha_j$), aminoacid propensities ($\Omega_j$), *etc.* We set constant weights $w_j = 1$ (reduction to adjacency) or equal to node degree $w_j = \delta_j$ when we do not know the properties of nodes.

Step 4 - MI transforms **W** into a Markov Matrix ${}^{1}\mathbf{\Pi}$ and obtain the natural powers of this matrix ${}^{k}\mathbf{\Pi} = ({}^{1}\mathbf{\Pi})^{k}$. According to Markov Chains theory the elements of these matrices ${}^{k}p_{ij}$ are probabilities of short/long-range interactions for pairs of nodes place at topological distances $d_{ij} \leq k$.

Step 5 - MI use the values of ${}^{k}\mathbf{\Pi}$ matrices to calculate different molecular descriptors ${}^{k}D_t(G_m)$ for small molecules. The classic MI can be used for small molecules (drugs, metabolites, etc.) or biopolymers (proteins, RNAs, DNA). MI-NODES is used to read the files with the structure of complex networks.

## 3. MI PARAMETERS

### 3.1. MI parameter for drugs

MI calculate different types of molecular properties ${}^{k}D_t(G_m, w_j)$ [92, 95, 100] based on the molecular graph $G_m$ of the $m^{th}$ molecule and weights of nodes (atoms) equal to physicochemical atomic properties ($w_j$).

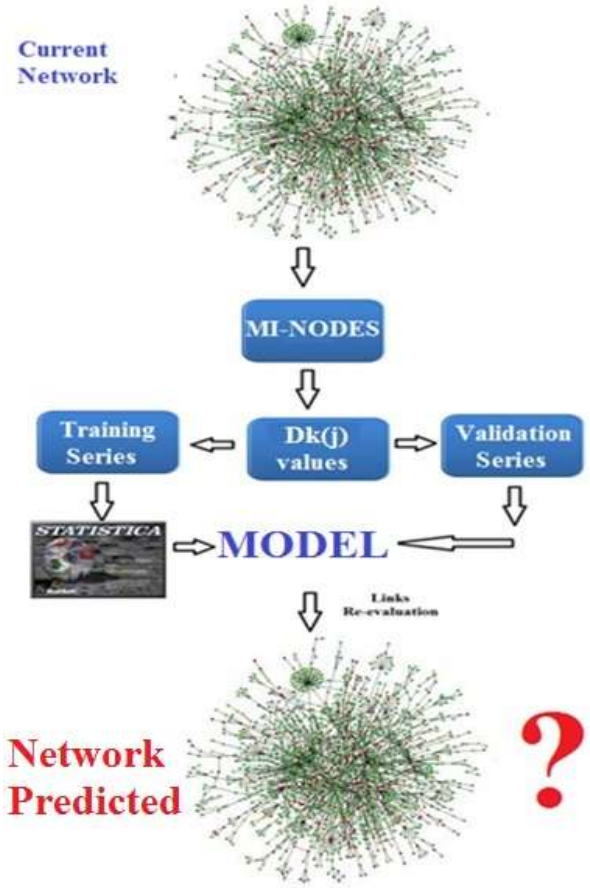



**Figure 2.** QSPR analysis of complex networks

We can omit $w_j$ in the notation when we use only one atomic property and declare it *a priori*, *e.g.* $w_j = \chi_j$ the atomic electronegativity. For instance, it is possible to calculate mean atomic electronegativities ${}^{k}D_{\chi}(G_m)$, Shannon entropy of electron delocalization ${}^{k}D_{\theta}(G_m)$, or spectral moments ${}^{k}D_{\pi}(G_m)$ [91, 101].

$$ {}^{k}D_{\chi}(G_m) = \sum_{j \in G} p_k(G_m) \cdot \chi_j \quad (1) $$

$$ {}^{k}D_{\theta}(G_m) = -\sum_{j \in G}^{n} {}^{k}p_j(G_m) \cdot \log\left[ {}^{k}p_j(G_m) \right] \quad (2) $$

$$ {}^{k}D_{\pi}(G_m) = \sum_{i=j \in R}^{n} {}^{k}p_{ij}(G_m) \quad (3) $$

It is possible to consider isolated atoms ($k = 0$) in a first estimation of the molecular properties ${}^{0}D_{\chi}(G_m)$, ${}^{0}D_{\theta}(G_m)$, or ${}^{0}D_{\pi}(G_m)$. In this case, the probabilities ${}^{0}p_{ij}(w_j)$ are determined without considering the formation of chemical bonds (simple additive scheme). It is possible to consider the gradual effects of the neighboring atoms placed at distance *k* using the absolute probabilities $p_k(w_j)$ with which these atoms affect the contribution of the atom *j* to the molecular property in question.

### 3.2. MI parameters for protein 3D structures

In the MI algorithm, we codify the information about protein structure using a Markov matrix ${}^{1}\mathbf{\Pi}$ that quantify the probabilities of short-term field interactions among amino acids (*aa*) [9, 72, 102-104]. The matrix ${}^{1}\mathbf{\Pi}$ is constructed as a squared matrix ($n \times n$), where *n* is the number of amino *aa* in the $m^{th}$ protein with contact map represented by the graph $G_m$ [105-107] In previous works we have predicted protein function based on mean values of 3D-Potentials ${}^{k}D_{\xi}(G_m, E)$, ${}^{k}D_{\xi}(G_m, vdW)$, and ${}^{k}D_{\xi}(G_m, h)$ for different type of interactions or molecular fields derived from ${}^{1}\mathbf{\Pi}$. The main types of the molecular fields used are: Electrostatic (e), van der Waals (vdw), and HINT (h) potentials [106, 108, 109]. The detailed explanation has been published before. In some of these works we calculated also entropy ${}^{k}D_{\xi}(G_m, E)$, ${}^{k}D_{\xi}(G_m, vdW)$, and ${}^{k}D_{\xi}(G_m, h)$ and spectral moment ${}^{k}D_{\theta}(G_m, E)$, ${}^{k}D_{\theta}(G_m, vdW)$, and ${}^{k}D_{\theta}(G_m, h)$ values for the same molecular fields. See the formula of the 3D mean potential, entropy, and moments for the electrostatic field:

$$ {}^{k}D_{\xi}(G_m) = -\sum_{j \in G_i} {}^{k}p_j(G_m) \cdot \xi_0(j) \quad (4) $$

$$ {}^{k}D_{\theta}(G_m) = -\sum_{j \in G}^{n} {}^{k}p_j(G_m) \cdot \log\left[ {}^{k}p_j(G_m) \right] \quad (5) $$

$$ {}^{k}D_{\pi}(G_m) = \sum_{i=j \in G}^{n} {}^{k}p_{ij}(G_m) \quad (6) $$

It is remarkable that the spectral moments depend on the probability ${}^{k}p_{ij}(G)$ with which the effect of the interaction f propagates from amino acid $i^{th}$ to other neighboring 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 ${}^{k}p_j(R)$ with which the amino acid $j^{th}$ has an interaction of type f with the rest of *aa*. The software MI [100] performs all these calculations by evaluation of the summation term either for all amino acids or only for some specific groups called regions ($R \in G_m$). We defined the regions in geometric terms and called them as core, inner, middle, or surface region. Please, see details in the literature [9, 72, 102-104, 109-113].

### 3.3. MI parameters for Complex Networks

In previous works, we have introduced new types of MI descriptors ${}^{k}D_t(G_m)$ complex networks. These values can be calculated as the sum of MI node centralities ${}^{k}C_t(j)$ for each $j^{th}$ nodes in the network, see **Table 1**. These descriptors are Markov chain generalizations of classic TIs. Some of these are Markov-Shannon Entropies [114], Markov-Randić indices [46], or Markov-Harary numbers [115]. We have used Markov-TIs to study several types of complex networks in Biology, Linguistics, Technology, Social, and Legal Sciences. In the next section, we describe different parameters of MI.

We implemented the new centralities in the software MI-NODES (MARCH-INSIDE for NOde DEScriptors) and used it to calculate the node centralities of the networks studied in this work. MI-NODES is a GUI Python/wxPython application developed by our groups. It is an upgrade of part of the code of the software MARCH-INSIDE adapted to manage any kind of complex networks. The program builds a Markov matrix (${}^{1}\mathbf{\Pi}$) for each network using as input the matrix of connectivity or adjacency of nodes often denoted as **A**. The elements of this stochastic matrix are the node-node transition probabilities ($p_{ij}$). The probability matrix is raised to the power *k*, resulting $({}^{1}\mathbf{\Pi})^{k}$. The resulting matrices ${}^{k}\mathbf{\Pi}$, wich are the $k^{th}$ natural powers of ${}^{1}\mathbf{\Pi}$, contain the transition probabilities ${}^{k}p_{ij}$. These are the probabilities to reach the $j^{th}$ node moving from the $i^{th}$ node throughout a walk of length k for each *k*. The generalization of the classic TIs and node centralities to general MI indices of order $k^{th}$ is straightforward to realize simply by substitution/multiplication of some parameters used in classic TIs like topological distances ($d_{ij}$) or node degrees (j) by/with Markov matrix parameters like transition probabilities ${}^{k}p_{ij}$. We can obtain different MI generalizations of classic TIs and/or node centralities. For instance, we can calculate k values of the new Markov-Rücker indices $WC_k(G)$ for a graph G (or probabilistic walk counts). We only have to change $d_{ij}$ by ${}^{k}p_{ij}$. Conversely, we can obtain k values of new Markov-Wiener indices $W_k(G)$ for a graph G multiplying $d_{ij}$ by ${}^{k}p_{ij}$. In so doing, we can run the sum over all nodes in G to calculate global TIs or only over all the $j^{th}$ nodes linked to one specific node i. The number of these nodes linked directly to one specific node is equal to $\delta_i$ (the degree of i) and we symbolized here a direct link as $j \rightarrow i$ and. In a very simple example, we can obtain a total of k values of new Markov-Rücker or probabilistic walk count centralities $WC_k(i)$ for the node $i^{th}$.

$$ {}^{k}D_{wc}(G) = \frac{1}{2}\cdot\sum_{i=1}^{D}\sum_{j=i}^{D} {}^{k}p_{ij} \qquad (7) $$

or

$$ {}^{k}D_{w}(G) = \frac{1}{2}\cdot\sum_{i=1}^{D}\sum_{j=i}^{D} {}^{k}p_{ij}\cdot d_{ij} $$

$$ {}^{k}C_{wc}(j) = \frac{1}{2}\cdot\sum_{i=1}^{1}\sum_{j\rightarrow i}^{\delta_i} {}^{k}p_{ij} = \frac{1}{2}\cdot\sum_{j\rightarrow i}^{\delta_i} {}^{k}p_{ij} \qquad (8) $$

or

$$ {}^{k}C_{w}(j) = \frac{1}{2}\cdot\sum_{i=1}^{1}\sum_{j\rightarrow i}^{\delta_i} {}^{k}p_{ij}\cdot d_{ij} = \frac{1}{2}\cdot\sum_{j\rightarrow i}^{\delta_i} {}^{k}p_{ij}\cdot d_{ij} $$

In **Table 1** we list the names, formula, software used for calculation, and references of many classic and MI centralities [33, 34, 46, 47, 49-52, 116].

## 4. GENERAL MI MODELS

### 4.1. Models of Drug-Target Networks (DT-Nets)

In MI strategy we can use as inputs the parameters of the $m^{th}$ drug molecule or protein ligands with molecular graph ($G_m = L_r$). We use $w_j = \chi_j$ by default, omit it in notations, obtaining the molecular descriptors ${}^{k}D_{\chi}(L_r)$, ${}^{k}D_{\theta}(L_r)$, or ${}^{k}D_{\pi}(L_r)$; by one hand. In addition, we should use the MI parameters of the $s^{th}$ protein sequence or 3D structure to obtain the descriptors ${}^{k}D_{\chi}(P_s)$, ${}^{k}D_{\theta}(P_s)$, or ${}^{k}D_{\pi}(P_s)$, by the other hand. We use the electrostatic field by default and omit it in notations. In the next lines we show the linear MI models for Drug-Protein Interactions (DPIs).

$$ S(DPI_{rs})_{pred} = \sum_{k=0}^{5} a_k \cdot {}^{k}D_t(L_r) + \sum_{k=0}^{5} b_k \cdot {}^{k}D_t(P_s) + c_0 \qquad (9) $$

The model deals with the calculation of score values (S) to predict the propensity of a set of compounds, to interact ($L_{rs} = 1$) or not ($L_{rs} = 0$) with different protein targets. A dummy input variable Affinity Class (AC) codify the affinity; AC = 1 for well-known DPIs and AC = 0 otherwise. This variable indicates either high (AC = 1) or low (AC = 0) affinity of the $r^{th}$ drug or protein by the $s^{th}$ target protein. The parameter $S(DPI_{rs})_{pred}$ is the output of the model and it is a continuous and

dimensionless score that give higher values for DPIs and lower values for nDPIs. In the model, $a_k$, $b_k$, $c_k$, and $d_k$ represents the coefficients of the MI function determined using the software STATISTICA 6.0 software package [117]. In all these cases, as well as in all the following models presented here, we can check the Specificity (Sp), Sensitivity (Sn), total Accuracy (Ac), or the Area Under the ROC curve (AUROC) to determine the goodness-of-fit to data in training and external validation series.

### 4.2. MI models of Complex Networks (Nets)

We can seek a linear function able to discriminate between two classes of pairs of nodes, linked and not linked in a new model network. The data necessary to train the model are obtained from the different systems studied and include two types of pairs of nodes (categorical dependent variable): linked ($L_{ij}$ = 1) and not linked ($L_{ij}$ = 0). The MI function has the following form:

$$S(L_{ij}) = \sum_{k=0}^{5} a_{ik} \cdot {}^{k}C_t(i) + \sum_{k=0}^{5} b_{jk} \cdot {}^{k}C_t(j) + \sum_{k=0}^{5} c_{ijk} \cdot \left[ {}^{k}C_t(i) - {}^{k}C_t(j) \right] + \sum_{k=0}^{5} d_{ijk} \cdot {}^{k}C_t(i) \cdot {}^{k}C_t(j) + e_0 \quad (10)$$

The continuous dependent variables used are: the node centralities of order k and type t for the two nodes ${}^{k}C_t(i)$, ${}^{k}C_t(j)$ and functions of these node centralities like $[{}^{k}C_t(i) - {}^{k}C_t(j)]$ and ${}^{k}C_t(i) \cdot {}^{k}C_t(j)$. Here we use the symbol ${}^{k}C_t$ instead of ${}^{k}D_t$ (the symbol used in the previous examples). This difference indicates that in the previous examples of MI models we talk in general about descriptors ${}^{k}D_t$ (centralities or not) of a molecular graph. However, in this example we are talking about node centralities ${}^{k}C_t$. Therefore we have Nv = 4·k·t variables that encode information of the pair of nodes ij and its neighbors (placed at a topological distance d = k). The parameters $a_{ik}$, $b_{jk}$, $c_{ijk}$, and $d_{ijk}$ are coefficients for variables and $a_0$ the independent term. $S(L_{ij})$ is the output variable (a real number).

### 4.3. Models with MA operators

Let be $S_j$ the output variable of a model used to score the quality of the connectivity pattern $L_{ij}$ between the node $i^{th}$ and all the remnant ($n$ -1) nodes in the network. In this sense, $S_j$ is a real valued variable that scores the quality of the connectivity pattern or links (all direct and indirect connections) established between the node $j^{th}$ and the other nodes. The higher is the value of $S_j$ the closer to the correct pattern are the links set for $j^{th}$ in the network as a whole, according to the model. On the other hand, $L_j$ is the input dependent variable. $L_j$ = 1 when a node is correctly linked to the rest of the nodes in the network and $L_j$ = 0 when a node has a random connectivity model. We can use linear algorithm like Linear Discriminant Analysis (LDA) or a Linear Neural Network (LNN) to fit the coefficients $a_k$, ${}^{g}b_k$, and $c_0$. We can use also a non-linear methods, e.g., Artificial Neural Networks (ANNs) [118]. The linear equation case is:

$$S_j = \sum_{k=0}^{5} a_k \cdot {}^{k}C_t(j) + \sum_{g=0}^{g=Ng} \sum_{k=0}^{5} b_{gk} \cdot \left[ {}^{k}C_t(j) - {}^{k}C_t(j)_{g-avg} \right] + c_0 = \sum_{k=0}^{5} a_k \cdot {}^{k}C_t(j) + \sum_{g=0}^{g=Ng} \sum_{k=0}^{5} b_{gk} \cdot \Delta {}^{k}C_t(j)_g + c_0 \quad (11)$$

In this equation we can see the coefficients ($a_k$) of the Wiener-Markov centralities used as input $W_k(j)$ and/or the coefficients (${}^{g}b_k$) of different deviation terms constructed with these variables. The deviation terms have the general form $\Delta W_k(j)_g = [W_k(j) - W_k(j)_{g.avg}]$. Where, $W_k(j)_{g.avg}$ is the average value (avg) of $W_k(j)$ for a sub-set or group (g) of nodes of the same graph G ($g \in G$) that obey a given condition. This type of deviation terms resembles the moving average terms used in time series models like in Box–Jenkins' ARIMA models [45]. However, in the present work g may be not only a time frame or season (laws approved in the same year) but also a biological boundary (metabolic reactions in the same organism) or spatial condition (interactions in the same eco-system); see results section.

## 5. EXAMPLES OF MI MODELS

### 5.1. Markov-Shannon Entropy models

Entropy measures are universal parameters useful to codify biologically-relevant information in many systems. Kier published probably the first work on the use of Shannon's entropy to calculate an structural information parameter (called molecular negentropy) and carry out QSPR studies [119, 120]. Graham *et al.*[121-126] used entropy measures to study the information properties of organic molecules. In any case, Shannon's entropy have been used to describe not only small molecules [120, 127-134] but also protein [135, 136] or DNA sequences [137] as well as protein interaction

networks [138]. Mikoláš *et al.* [139] reviewed the use of entropy measures in functional magnetic resonance (fMRI). The software MI calculates values of Markov-Shannon entropy for both molecular structures (drugs and target proteins) and nodes centralities in complex networks [84, 92]. Last year [47], we published a paper on the QSPR study of complex molecular systems and social networks using entropy measures and one alignment-free, multi-target, and multi-scale algorithm (see **Figure 1)**. The procedure is essentially the same than in classic QSPR studies with some variations in each problem. In the following sections, we review some of these MI models for illustrative purposes. The first model was developed to predict the DT-Net of FDA approved drugs. The prediction of DT-Nets is important due to the high cost of the experimental [78, 140, 141]. Here, we have developed a model that takes into account the structure of the drug, the structure of the target, and the information about the drug/target nodes in the studied network (see **Figure 3**).

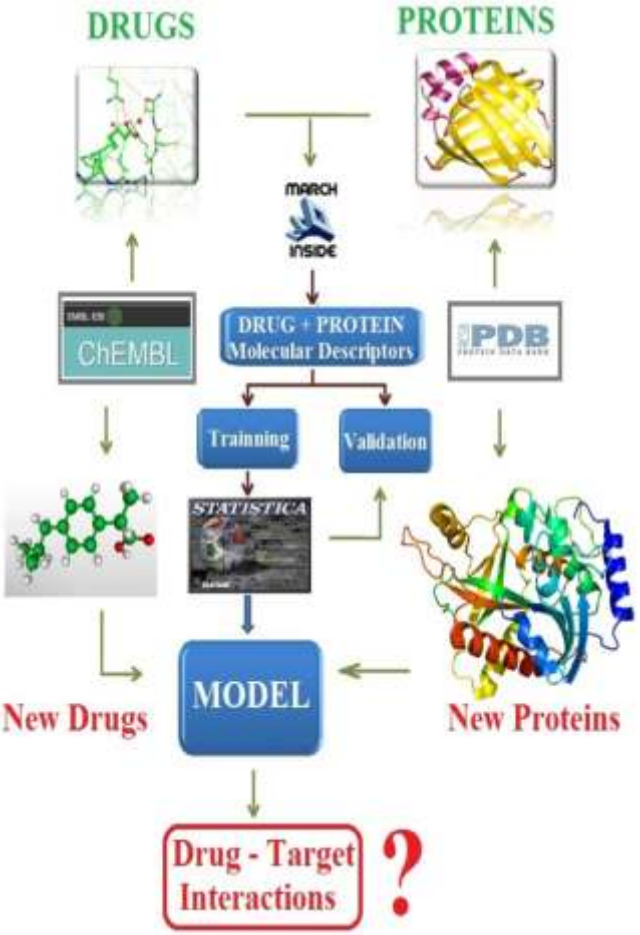



**Figure 3.** QSPR analysis of drug-target networks

In this network $L_{rs}$ = 1 if the $r^{th}$ protein ($P_r$) is a target of the $s^{th}$ drug or ligand ($L_s$) in the DrugBank database and $L_{rs}$ = 0 otherwise. The best model found was:

$$S(L_{rs}) = +0.11 \cdot {}^{0}D_{\theta}(L_r) - 0.47 \cdot {}^{4}D_{\theta}({}^{m}P_s) - 2.19 \cdot {}^{3}C_{\theta}(j)_P - 1.10 \cdot {}^{5}C_{\theta}(j)_L - 1.43 \quad (12)$$

$$n = 2{,}234 \quad \chi^2 = 2{,}123 \quad p < 0.001$$

Where, ${}^{k}D_{\theta}(L_r)$ and ${}^{k}D_{\theta}({}^{m}P_s)$ are the Markov-Shannon entropy descriptors used to codify the information about the structur of the drug and the protein. Specifically the descriptor of the protein includes amino-acids placed only in the middle region (m) of the target proteins (see details about protein descriptors in the previous sections).

In addition, $C_{\theta}(j)_L$ and ${}^{k}C_{\theta}(j)_P$ are centralities of the nodes for the drug/ligand and the target in the DT-Net. This put in evidence the multiscale nature of the model with descriptors for drugs, proteins, and nodes in the DT-Net. The $\chi^2$ = 2,123 statistics corresponds to a p-level < 0.001, which indicates a significant discrimination ratio. The values of Ac, Sn, and Sp were very good for validation and training series, see details in the reference [47].

On the other hand, the study of Metabolic Reaction Networks (MR-Nets) is of great interest in biology because many applications are directly built on the use of cellular metabolism in Biotechnology and Biomedicine [142, 143]. In this sense, computational studies of MR-Nets become very useful [144, 145].

In a recent work [47], we developed a model to predict the correct connectivity patterns in MRNs using as inputs the Markov-Shannon entropy centralities ${}^{k}C_{\theta}(j)$ for nodes in already-known networks. For this analysis, we have used metabolic networks of four model organisms belonging to different domains of the tree of life. These organisms are *Escherichia coli* (EC), *Saccharomyces cerevisiae* (SC), *Caenorhabditis elegans* (CE), and *Oryza sativa* (OS). They cover important branches of the tree of live including a gram-negative bacterium [146-157], a fungus with industrial importance [158], free-living nematode that has become a popular model for genetic [159-162], and the most widely studied model for cereals [163], respectively.

The best MI-Entropy model found was:

$$S(L_{ij}) = 159.16 \cdot {}^{3}C_{\theta}(e_i) - 120.70 \cdot {}^{1}C_{\theta}(p_j) - 95.42 \cdot \left[{}^{5}C_{\theta}(e_i) - {}^{5}C_{\theta}(p_j)\right] - 0.26 \quad (13)$$

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

In this equation, $S(L_{ij})$ is a real-valued output variable that scores the propensity of the $i^{th}$ input or educt ($e_i$) (reactant or substrate) to undergo a metabolic transformation into the product ($p_j$). The entropy parameters quantify the information related to middle-long range subsequent metabolic transformations of all the neighbors of the input-output metabolites (k = 5) in the metabolic network. See results in **Table 2**.

Also, the importance for the human and animal health and therefore for the economy, much attention has been focused on complex network analysis of parasite-host interactions [164]. However, the high experimental difficulty inherent to the *in situ* determination these interactions make the use a computational model a very interesting option.

In this work, we used ${}^{k}C_{\theta}(j)$ values to seek a QSPR-like model able to predict HP-Nets. The best model found for the HP-Net was:

$$S(L_{ij}) = -82.62 \cdot \left[{}^{5}C_{\theta}(p_i) - {}^{5}C_{\theta}(h_j)\right] - 5.52 \quad (14)$$

$$n = 49{,}218 \quad \chi^2 = 21{,}728 \quad p < 0.001$$

In this equation, $S(L_{ij})$ is a real-valued output variable that scores the propensity of the $i^{th}$ parasite specie ($p_i$) to infect a given host specie ($h_j$). See results in **Table 2**. Connectivity is also the key to understanding distributed and cooperative brain functions and can be represented by Brain Connectome Networks (BC-Nets) [165].

The eventual impact and success of connectivity databases, however, will require the resolution of several methodological problems that currently limit their use. These problems comprise four main points: (i) objective representation of coordinate-free, parcellation-based data, (ii) assessment of the reliability and precision of individual data, especially in the presence of contradictory reports, (iii) data mining and integration of large sets of partially redundant and contradictory data, and (iv) automatic and reproducible transformation of data between incongruent brain maps [166].

In order to address points (ii) and (iv), we have developed a specific model for the 'collation of connectivity data on the macaque brain' (CoCoMac) database (http://www.cocomac.org). The best model found for this BC-Net was:

$$S(L_{ij}) = 70.56 \cdot {}^{1}C_{\theta}(a) + 74.51 \cdot {}^{5}C_{\theta}(e) - 1.75 \quad (15)$$

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

In this equation, $S(L_{ij})$ is a real-valued output variable that scores the propensity of the $i^{th}$ cerebral cortex region to undergo co-activation with the $j^{th}$ region in the CoCoMac network. The entropy parameters quantify the information related to the position of the afferent/efferent regions and their direct neighbors (k = 1) in the network. The model showed very good results (see **Table 2)**.

Another important problem to be studied with networks is the spreading of diseases. For instance, Fasciolosis is a parasitic infection caused by *Fasciola hepatica* (liver fluke) that has become an important cause of lost productivity in livestock worldwide. It is considered a secondary zoonotic disease until the mid-1990s, human fasciolosis is at present emerging or re-emerging in many countries. In addition, it presents a range of epidemiological characteristics related to a wide diversity of environments [167].

In this sense, the study of geographical spreading of fasciolosis becomes a subject of great interest. In fact, in a recent work we have constructed a Fasciolosis Epidemiology network (FE-Net) to study the landscape spreading of fasciolosis in Galicia (NW Spain) [168]. However, we do not have quantitative criteria on the quality of the network connectivity, and re-sampling of all data to re-evaluate this connectivity in a field study is a hard and expensive task in terms of time and resources.

This situation has prompted us to seek a model in order to assess the quality of the network previously assembled. The best QSPR model found and published in our previous work for the FE-Net was:

$$S(L_{ij}) = -20.23 \cdot {}^{1}C_{\theta}(f_i) + 165.13 \cdot {}^{4}C_{\theta}(f_j) - 0.82 \quad (16)$$

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

The entropy used in this equation quantify information about the connectivity patterns between farms in the network **C**.

**Table 2.** MI models of complex networks

| Net | Par. | $^{k}C_{BM}(j)$ | $^{k}C_{\theta}(j)^{k}_{j}$ | $^{k}C_{\pi}(j)^{k}_{j}$ | $^{k}C_{wc}(j)$ | $^{k}C_{\chi}(j)$ |
|---|---|---|---|---|---|---|
| Train | | | | | | |
| MR | Sp | 72.22 | 99.98 | ? | 81.32 | 70.19 |
| | Sn | 71.25 | 87.24 | ? | 73.91 | 70.63 |
| PH | Sp | 87.49 | 95.4 | 87.49 | 95.24 | 90.56 |
| | Sn | 100 | 72.22 | 100 | 73.27 | 92.70 |
| BC | Sp | 84.14 | 92.2 | 98.49 | 88.40 | 75.32 |
| | Sn | 72.70 | 71.2 | 73.30 | 74.64 | 94.69 |
| FE | Sp | 87.14 | 99.2 | 93.21 | 71.49 | 100 |
| | Sn | 72.68 | 70.4 | 72.01 | 71.64 | 89.70 |
| Validation | | | | | | |
| MR | Sp | 72.28 | 99.96 | ? | 81.82 | 71.17 |
| | Sn | 71.24 | 86.91 | ? | 73.81 | 70.89 |
| PH | Sp | 87.67 | 95.5 | 87.67 | 95.43 | 91.00 |
| | Sn | 100 | 72 | 100 | 70.81 | 92.83 |
| BC | Sp | 84.42 | 92.5 | 98.41 | 88.30 | 75.51 |
| | Sn | 71.88 | 70.4 | 71.21 | 73.27 | 94.73 |
| FE | Sp | 87.34 | 99.1 | 93.20 | 71.55 | 100 |
| | Sn | 75.78 | 74.2 | 73.47 | 70.54 | 90.22 |
| Mod. | Ref. | [52] | [47] | [116] | [51] | |

Net = Network: 1- Metabolic Reactions Network (MR-Net), 2 - Parasite-Host Net (PH-Net), 3 - Brain Connectome Net (BC-Net), 4 - Fasciolosis Epidemiology Net (FE-Net). Par. = Parameter: Sp = Specificity and Sn = Sensitivity. Ref. = Reference where the model was published.

As can be seen in the equations described in materials and methods, the connectivity of **C** depends on the spatial coordinates ($x_i$, $y_i$) of the farm ($f_i$), the altitude of the place ($h_i$), and the anti-parasite drug treatment ($Tr_j$) used to prevent Fasciolosis in this farm. Consequently the matrix **C** quantifies the *a priori* propensity $C_{ij}$ = 1 of this disease to spread between farms immediately after treatment depending on geographical conditions.

On the other hand, matrix **L** includes both criteria: (i) the preexistence of a high propensity for disease spreading $C_{ij}$ = 1 and (ii) the experimental confirmation $L_{ij}$ = 1 of a high Risk Ratio ($RR_{ij}$) of Prevalence After Treatment ($PAT_j$) for this disease in farms. See **Figure 4**, published before in one of our papers [52], see also the section about auto-correlation indices. The QSPR equation developed here was obtained by studying **L** and the model presents good values of Sensitivity (Sn), and Specificity (Sp), see **Table 2.**

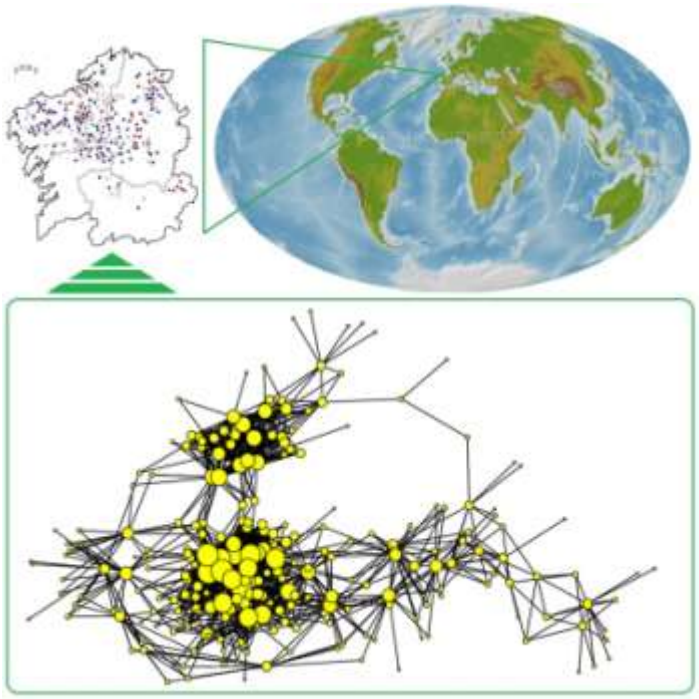


**Figure 4.** Top left: Geographical map of Galicia (NW Spain) showing the location of the 275 sampled farms: the status of infection (empty circles: *F. hepatica* free and filled circles: *F. hepatica* infected) and the treatment administered on each farm are shown (blue: none; red: anthelmintic effective against fluke mature stages and green: a fasciolicide effective against immature and mature stages). Bottom: Fasciolosis landscape-spreading network. The size of each node represents its degree.

Another MI-Shannon entropy model published in the previous work is useful to study the SL-Net for Spain's law system. The use of network analysis methods in social sciences began in 1930 and today are widely used [169]. However, the application of these methods in legal studies is still at the beginning [170-172]. Network tools may illustrate the interrelation between the different law types and help to understand law consequences in society and its effectiveness or not. We have used the list of the financial laws to construct the network described. The best model found was:

$$S(L_{ij}) = 650.88 \cdot \left[ {}^{1}C_{\theta}\left({}^{c}L_{ti}\right) - {}^{1}C_{\theta}\left({}^{c}L_{ti+1}\right) \right] + 0.12 \quad (17)$$

$$n = 33{,}951 \quad \chi^2 = 32{,}942 \quad p < 0.001$$

Where the two parameters in the equation are the entropy parameters that quantify information about the Legal norms (Laws) of type L introduced in the Spanish legal system at time $t_i$ and $t_{i+1}$ with respect to the previous or successive $k^{th}$ norms approved.

The model behaves like a time series embedded within a complex network. This is because it predicts the recurrence of the Spanish law system to a financial norm of class c when socio-economical conditions change at time $t_{i+1}$ given that have been used a known class of norm in the past at time $t_i$. The model correctly re-constructed the network of the historic record for the Spanish financial system with high Sp and Sn **(Table 2)**. In **Figure 5,** we illustrate the steps used to develop the MI model of this network; which is also a hierarchical time series.


**Figure 5.** QSPR analysis of one SL-Net

The last MI-Shannon entropy model reported is useful to predict the Network (WT-Net) of Smart Package for World's food industry. Traditionally, the basic functions of packaging have been classified into 4 categories: protection, communication, convenience, and containment [173]. Smart or Active Packaging is an innovative concept that can be defined as a mode of packaging in which the package, the product, and the environment interact to prolong shelf life or enhance safety or sensory properties, while maintaining the quality of the product [174]. In addition there is a growing concern about foodborne diseases, and many companies are interested in the development of biosensors included in the packages in order to detect the presence of pathogens [173]. In the previous work we studied a large world-trading network (WT-Net) for the current world trade (year 2011) of smart packaging for food industry, interconnecting categories like Country (CU), Company (CO), Product (PR), Food Type (FT), and product use or Packaging Type (PT), see also datasets section. The best model found was:

$$S(L_{ij}) = -2.00 \cdot {}^{1}C_{\theta}(i) - 142.87 \cdot {}^{1}C_{\theta}(j) + 116.65 \cdot {}^{5}C_{\theta}(j) + 0.72 \quad (18)$$

$$n = 31{,}911 \quad \chi^2 = 19{,}022 \quad p < 0.001$$

The model presents very good values of Sn and Sp (see **Table 2**). The first parameter quantifies the information referred to the trading relationships of the $i^{th}$ node with its direct neighbors (k = 1) in the world trade network. The second parameter quantifies the same information for the $j^{th}$ node and its direct neighbors (k = 1). The last parameter quantifies the information referred to middle-long range trading relationships (k = 5) in the trade network between the $j^{th}$ node and its neighbors of any class. In order to use this equation, it is necessary to introduce the values of the centralities for the $i^{th}$ and $j^{th}$ nodes according to the following hierarchical order in i to j direction: Country (CU) → Company (CO) → Product (PR) → Packaging Type (PT) → Food Type (FT) if we want to predict the expected success of a given CO to introduce a determined PT in the WT-Net. In **Figure 6,** we illustrate the network for a better understanding.

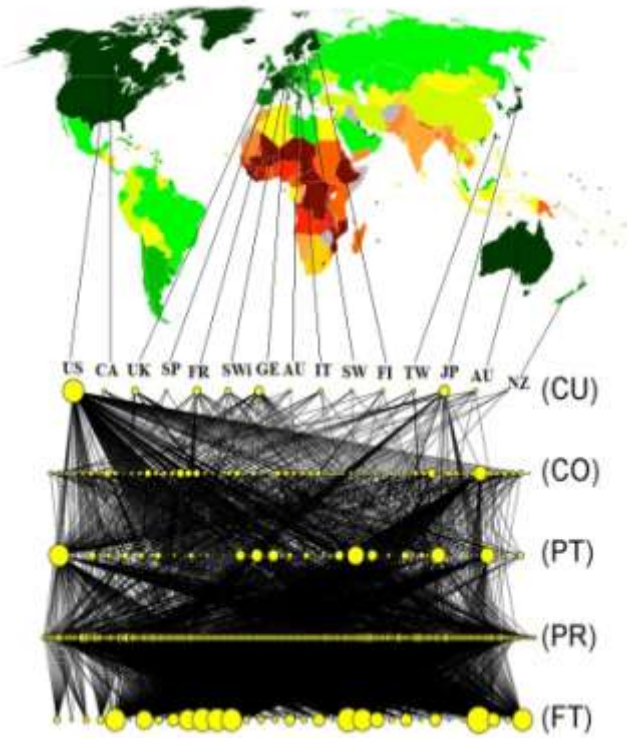



**Figure 6.** WT-Net of Smart Packaging for Food Industry.

### 5.2. Rücker-Markov centralities models

Rücker and Rücker [175, 176] published a series of works about the use of Walk Count (WC) indices, in this sense. In this previous work, it is demonstrated how the complexity of a (molecular) graph can be quantified in terms of the walk counts, extremely easily obtained graph invariants that depend on size, branching, cyclicity, and edge and vertex weights (w). Weights are important to numerically quantify properties that differentiate classes of nodes. Gutman co-authored another paper with Rücker & Rücker about WCs [177]. They reviewed applications of WCs in theoretical chemistry based on the fact that the (i, j)-entry of the $k^{th}$ power of the adjacency matrix is equal to the number of walks starting at vertex i, ending at vertex j, and having length k. In 2003, the concept was extended by Lukovits and Trinajstić [178] to zero and negative orders. More recently, Bonchev has applied WCs and other TIs to the complexity analysis of yeast proteome network [3]. In a recent work, we introduced the new Rücker-Markov indices ${}^{k}C_{wc}(j)$ [179] and use them to seek QSPR models able to predict of the connectivity of new complex networks. For instance, we used ${}^{k}C_{wc}(j)$ values to seek a QSPR-like model able to predict PH-Nets, the DS-Net of Fasciolosis in Galicia, and the BC-Net reported in CoCoMac experiment. The best models found for each one of these datasets were the following, in this order:

$$S(L_{ij}) = -258.93 \cdot \left[ {}^{1}C_{wc}(p_i) - {}^{1}C_{wc}(h_j) \right] + 283.69 \cdot \left[ {}^{2}C_{wc}(p_i) - {}^{2}C_{wc2}(h_j) \right] - 88.75 \cdot \left[ {}^{4}C_{wc}(p_i) - {}^{4}C_{wc}(h_j) \right] + 0.25 \quad (19)$$

$$n = 49{,}218 \quad \chi^2 = 22{,}297 \quad p < 0.001$$

$$S(L_{ij}) = 8.34 \cdot \left[ {}^{1}C_{wc}(f_i) - {}^{1}C_{wc}(f_j) \right] - 2.17 \cdot \left[ {}^{5}C_{wc}(f_i) - {}^{5}C_{wc}(f_j) \right] - 0.56) \quad (20)$$

$$n = 23{,}991 \quad \chi^2 = 1{,}965 \quad p < 0.001$$

$$S(L_{ij}) = 1.92 \cdot {}^{1}C_{wc}(i) + 2.14 \cdot {}^{2}C_{wc}(j) - 1.68 \quad (21)$$

$$n = 39{,}070 \quad \chi^2 = 20{,}602 \quad p < 0.001$$

In these equations, $S(L_{ij})$ is a real-valued output variable that scores the propensities with which the $i^{th}$ parasite specie ($p_i$) infect host specie ($h_j$), the disease spreads from the $i^{th}$ farm to the $j^{th}$, or the the $i^{th}$ cerebral region to co-activate with the $j^{th}$ region. You can compare the results for those and other models in **Table** 2.

### 5.3. Broto-Moreau stochastic centralities models

In the 1980s, Broto & Moreau applied an autocorrelation function to the molecular graph in order to measure the distribution of atomic properties on the molecular topology. This measure was called Autocorrelation of Topological Structure (ATS) or Broto-Moreau autocorrelation indices (BMis) [180-182]. The idea of ATS has been re-formulated in different ways in order to incorporate more information. Moro studied electrostatic potential surface properties [183], Caballero and Fernández [184-187] carry out QSPR in proteins. Some ATS models have been implemented in web servers such as IUPforest-L [188] and PROFEAT [189]. We implemented them in the software S2SNet (Sequence to Star Networks) [190], to calculate ATS indices for mass spectra signals of proteins, 1D NMR signals, IR spectra, time series data, texts and any other type of string data. In a recent work we studied similar datasets than in the two previous examples but using the MI autocorrelation centrality values ${}^{k}C_{BM}(j)$ [191]. The best model for the MR-Nets of the organisms EC, SC, CE, and OS, PH-Nets, BC-Net of macaque

visual cortex, and DS-Net for Fasciolosis in Galicia are the following, see also **Table 2**.

$$S(L_{ij}) = -0.73 + 23.44 \cdot {}^{5}C_{BM}(e_i) - 5.59 \cdot \left[{}^{3}C_{BM}(e_i) - {}^{3}C_{BM}(p_j)\right] \quad (22)$$

$$n = 74{,}999 \quad \chi^2 = 20{,}143 \quad p < 0.001$$

$$S(L_{ij}) = 4.59 \cdot \left[{}^{2}C_{BM}(p_i) - {}^{2}C_{BM}(h_j)\right] + 0.21 \quad (23)$$

$$n = 49{,}218 \quad \chi^2 = 15{,}801 \quad p < 0.001$$

$$S(L_{ij}) = 12.74 \cdot \left[{}^{1}C_{BM}(i) - {}^{1}C_{BM}(j)\right] - 0.80 \quad (24)$$

$$n = 24{,}956 \quad \chi^2 = 9{,}422 \quad p < 0.001$$

$$S(L_{ij}) = -0.07 - 11.50 \cdot {}^{3}C_{BM}(f_i) - 18.26 \cdot \left[{}^{1}C_{BM}(f_i) - {}^{1}C_{BM}(f_j)\right] \quad (25)$$

$$n = 23{,}377 \quad \chi^2 = 3{,}897 \quad p < 0.001$$

### 5.4. Wiener-Markov centralities models

In 1947, Wiener published an article entitled *Structural determination of paraffin boiling points* [192]. In this work it is proposed that organic compounds, as well as all their physical properties, depend functionally upon the number, kind and structural arrangement of the atoms in the molecule [193-195]. Hosoya coined one term of Wiener's equation in 1971 as the Z index [196-198].

The Wiener index (W) index was independently proposed in 1959 by Harary in the context of sociometry, with the name *total status of a graph* [199] as well as in 1975 by Rouvray and Crafford [200]. In any case, W index or path number is calculated as the half sum of all the elements $d_{ij}$ of the distance matrix (**D**). More distant atom pairs make larger contribution to W than adjacent atom pairs:

$$W = \frac{1}{2} \cdot \sum_{i=1}^{D} \sum_{j=1}^{D} d_{ij} \quad (26)$$

In a very recent work [53], we used Markov-Wiener centralities ${}^{k}C_w(j)$ to predict correct connectivity patterns of nodes in MR-Nets of 43 organisms using MIANN models (acronym formed by MI and ANN)[9]. In **Table 3** we depict the classic parameters and the average values of ${}^{k}C_w(j)$ for the full MR-Nets of many organisms. These average values are the inputs used to characterize the organisms with the MI method in the predictive MIANN models. After that we tested different MIANN models using as inputs the values of ${}^{k}C_w(j)$ and with linear (LNN) and non-linear (ANN) topologies in of the ANN.

In **Table 4**, we can see that the best MIANN model found presents very good values of Accuracy, Sensitivity, and Specificity for the recognition of links both in training and external validation series. The models were obtained using as input 15 descriptors: 5 Markov-Wiener centralities ${}^{k}C_w(j)$, 5 MA values denoted as ${}^{k}C_w(j)_{g.avg}$ and 5 deviation terms $\Delta\ {}^{k}C_w(j)_g$. Multilayer Perceptron (MLP) [201] method fails to generate good prediction models, since it presents values of Specificity and Sensitivity close to 50%. On the other hand, the LNN based on 15 descriptors (LNN 15:15-1:1) is able to classify correctly a 78.1% of the cases, with a sensitivity of 77.9% and a specificity of 77.6%. The LNN is equivalent to a LDA equation, the simplest type of classification model.

We also developed a MIANN-Wiener models of BI-Nets published in IWDB. The results are presented in **Table 4**. We obtained the best classification model for IWDB with the MLP classifier based on 13 input descriptors and 13 neurons in the hidden layer (MLP 13:13-13-1:1). This model can classify 91.1% of the nodes with a sensitivity of 90.5% and specificity of 88.8%. Unlike the case of the MRNs, the LNN is not able to classify the nodes in the BI-Net with accuracy (<67%). Thus, it can be observed that, compared with the MR-Nets, the BI-Nets contain more complex information for the classification of the connectivity between nodes. The IWDBNs need complex classifiers such as MLPs in comparison with the MR-Nets that can be processed by using the simpler LNNs.

The **Figure 7** depicts one illustration of the IWDB. Last, we reported a MIANN-Wiener model for SL-Net of Spain's financial law system. These MIANN models behave like time series embedded within a complex network. The model predicts the recurrence of the Spanish law system to a financial norm of class c when the socio-economical conditions change at time $t_{i+1}$ given that have been used a known class of norm in the past at time $t_i$. The best model correctly re-constructed the network of the historic record for the Spanish financial system with high Sp and Sn (see **Table 4**).

**Table 3.** Classic parameters *vs.* Average values $^{k}C_{W}(j)_{org.avg}$ of metabolic networks of differentt organisms

| Organism | Classic parameters of MRNs | | | | | | | | Markov-Wiener centralities $^{k}C_{W}(j)_{org.avg}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| symbol | N | $L_{in}$ | $L_{out}$ | R | E | $g_{in}$ | $g_{out}$ | D | k = 1 | k = 2 | k = 3 | k = 4 | k = 5 |
| AA | 419 | 1278 | 1249 | 401 | 285 | 2.1 | 2.2 | 3.3 | 0.87 | 1.08 | 1.26 | 1.44 | 1.57 |
| AB | 395 | 1202 | 1166 | 380 | 271 | 2.1 | 2.2 | 3.2 | 0.88 | 1.07 | 1.24 | 1.43 | 1.56 |
| AG | 496 | 1527 | 1484 | 486 | 299 | 2.2 | 2.2 | 3.5 | 0.85 | 1.09 | 1.29 | 1.48 | 1.61 |
| AP | 204 | 588 | 575 | 178 | 135 | 2.2 | 2.2 | 3.2 | 0.95 | 1.11 | 1.25 | 1.46 | 1.6 |
| AT | 302 | 804 | 789 | 250 | 185 | 2.1 | 2.3 | 3.5 | 0.89 | 1.12 | 1.3 | 1.48 | 1.62 |
| BB | 187 | 442 | 438 | 140 | 106 | 2.3 | 2.4 | 3 | 0.8 | 0.99 | 1.18 | 1.37 | 1.49 |
| BS | 785 | 2794 | 2741 | 916 | 516 | 2.2 | 2.1 | 3.3 | 0.8 | 1.09 | 1.3 | 1.52 | 1.65 |
| CA | 494 | 1624 | 1578 | 511 | 344 | 2.1 | 2.2 | 3.3 | 0.83 | 1.08 | 1.28 | 1.46 | 1.59 |
| CE | 462 | 1446 | 1418 | 450 | 295 | 2.1 | 2.2 | 3.3 | 0.9 | 1.12 | 1.32 | 1.51 | 1.65 |
| CJ | 380 | 1142 | 1115 | 359 | 254 | 2.1 | 2.3 | 3.2 | 0.88 | 1.09 | 1.27 | 1.45 | 1.58 |
| CL | 389 | 1097 | 1062 | 333 | 231 | 2.1 | 2.2 | 3.3 | 0.88 | 1.1 | 1.3 | 1.51 | 1.63 |
| CQ | 194 | 401 | 391 | 134 | 84 | 2.2 | 2.3 | 3.4 | 0.99 | 1.14 | 1.27 | 1.47 | 1.62 |
| CT | 215 | 479 | 462 | 158 | 94 | 2.2 | 2.4 | 3.5 | 0.9 | 1.06 | 1.22 | 1.38 | 1.5 |
| CY | 546 | 1782 | 1746 | 570 | 370 | 2 | 2.2 | 3.3 | 0.88 | 1.13 | 1.33 | 1.56 | 1.68 |
| DR | 815 | 2870 | 2811 | 965 | 557 | 2.2 | 2.1 | 3.3 | 0.89 | 1.12 | 1.31 | 1.52 | 1.65 |
| EC | 778 | 2904 | 2859 | 968 | 570 | 2.2 | 2.1 | 3.2 | 0.79 | 1.03 | 1.24 | 1.44 | 1.57 |
| EF | 386 | 1244 | 1218 | 382 | 281 | 2.1 | 2.2 | 3.1 | 0.81 | 1.04 | 1.24 | 1.42 | 1.55 |
| EN | 383 | 1095 | 1081 | 339 | 254 | 2.1 | 2.2 | 3.3 | 0.89 | 1.11 | 1.31 | 1.5 | 1.65 |
| HI | 526 | 1773 | 1746 | 597 | 361 | 2.1 | 2.3 | 3.2 | 0.77 | 1.05 | 1.26 | 1.48 | 1.59 |
| HP | 375 | 1181 | 1144 | 375 | 246 | 2 | 2.3 | 3.3 | 0.89 | 1.11 | 1.3 | 1.5 | 1.62 |
| MB | 429 | 1247 | 1221 | 391 | 282 | 2.2 | 2.2 | 3.2 | 0.87 | 1.09 | 1.27 | 1.46 | 1.6 |
| MG | 209 | 535 | 525 | 196 | 85 | 2.4 | 2.2 | 3.5 | 0.96 | 1.14 | 1.26 | 1.38 | 1.48 |
| MJ | 424 | 1317 | 1272 | 415 | 264 | 2.2 | 2.3 | 3.5 | 0.88 | 1.11 | 1.29 | 1.47 | 1.6 |
| ML | 422 | 1271 | 1244 | 402 | 282 | 2.2 | 2.2 | 3.2 | 0.83 | 1.06 | 1.25 | 1.44 | 1.58 |
| MP | 178 | 470 | 466 | 154 | 88 | 2.3 | 2.2 | 3.2 | 0.91 | 1.11 | 1.29 | 1.46 | 1.59 |
| MT | 587 | 1862 | 1823 | 589 | 358 | 2 | 2.2 | 3.3 | 0.88 | 1.12 | 1.32 | 1.55 | 1.67 |
| NG | 406 | 1298 | 1270 | 413 | 285 | 2.1 | 2.2 | 3.2 | 0.85 | 1.06 | 1.24 | 1.42 | 1.56 |
| NM | 381 | 1212 | 1181 | 380 | 271 | 2.2 | 2.2 | 3.2 | 0.86 | 1.08 | 1.27 | 1.45 | 1.59 |
| OS | 292 | 763 | 751 | 238 | 178 | 2.1 | 2.3 | 3.5 | 0.93 | 1.19 | 1.39 | 1.57 | 1.71 |
| PA | 734 | 2453 | 2398 | 799 | 490 | 2.1 | 2.2 | 3.3 | 0.87 | 1.1 | 1.29 | 1.52 | 1.65 |
| PF | 316 | 901 | 867 | 283 | 191 | 2 | 2.3 | 3.4 | 0.93 | 1.14 | 1.33 | 1.5 | 1.65 |
| PG | 424 | 1192 | 1156 | 374 | 254 | 2.2 | 2.2 | 3.3 | 0.85 | 1.06 | 1.24 | 1.41 | 1.54 |
| PH | 323 | 914 | 882 | 288 | 196 | 2 | 2.2 | 3.4 | 0.92 | 1.12 | 1.31 | 1.49 | 1.63 |
| SC | 561 | 1934 | 1889 | 596 | 402 | 2 | 2.2 | 3.3 | 0.88 | 1.11 | 1.31 | 1.54 | 1.68 |
| ST | 403 | 1300 | 1277 | 404 | 280 | 2.1 | 2.2 | 3.1 | 0.89 | 1.08 | 1.24 | 1.44 | 1.57 |
| TH | 430 | 1374 | 1331 | 428 | 280 | 2.2 | 2.2 | 3.4 | 0.89 | 1.13 | 1.33 | 1.52 | 1.65 |
| TM | 338 | 1004 | 976 | 302 | 223 | 2.1 | 2.2 | 3.2 | 0.88 | 1.09 | 1.28 | 1.47 | 1.6 |

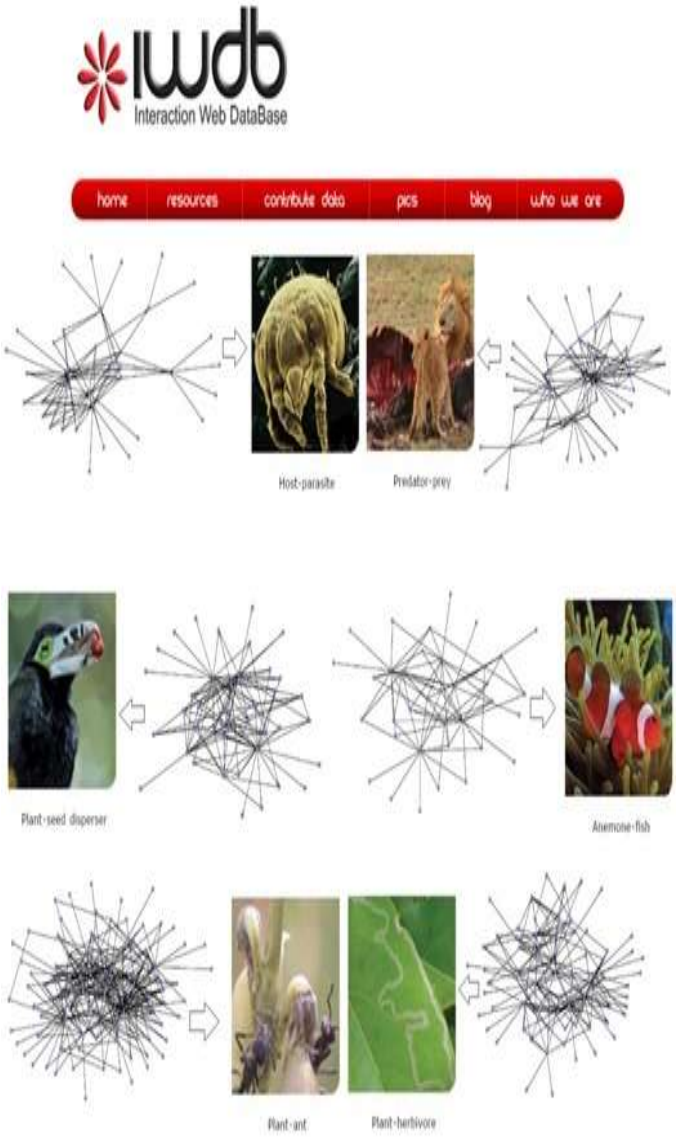



**Figure 7.** IWDB vs. BI-Nets

In this case, there is not a clear difference between the two models studied (LNN and MLP). In this situation, we can apply the Occam's razor and choose the LNN model, which is the simplest.

### 5.3. Markov-Balaban Index models

Prof. Alexandru T Balaban introduced one of the more famous TIs that have been widely-known as the Balaban's J index [202]. Blaban's J index have been used in many chemo-informatics to quantify structural information and include parameters like q = number of edges in the molecular graph, μ = (q − n + 1) = the cyclomatic number of the molecular graph, n = number of atoms in the molecular graph, and $S_i$ = distance sums calculated as the sums over the rows or columns of the topological distance matrix **D** of the graph G.

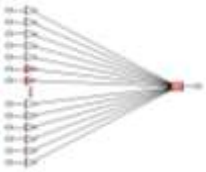
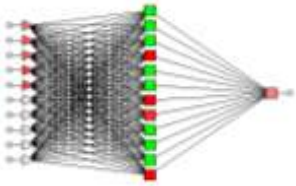
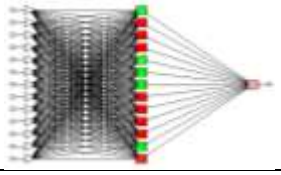
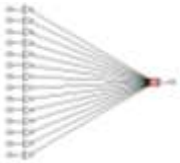
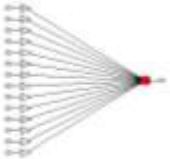
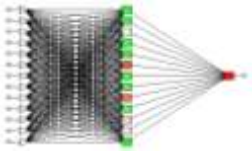
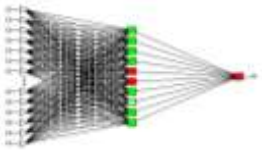
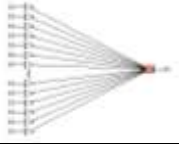
The formula of this classic TI is:

$$J(G) = \frac{q}{\mu + 1} \cdot \sum_{edges}^{q} \left(S_i \cdot S_j\right)^{-1/2} \qquad (27)$$

Many applications of Balban's J index deal with drug discovery; in particular the prediction of drugs with higher biological activity and/or low toxicity [203-210]. J index is useful as input for both linear and non-linear models like ANNs [211, 212]. J index have been used also to compare graphs or analyze combinatorial libraries and some authors have reported new generalizations of this index to create other TIs (called Balaban type parameters) [213-215].

For instance, Randić and Pompe[216]; reported the variable Balaban J index and the "reversed" Balaban index 1/J as well as a novel index 1/JJ derived from J and 1/J. In another very recent work [217], we introduced new Balaban type indices called the Markov-Balaban $^{k}C_J(j)$ centralities of order $k^{th}$ for the $j^{th}$ node in a complex network (see **Table 1**). In this previous work we also used multiscale MA operators to calculate deviation terms with the general form $\Delta\ ^{k}C_J(j)_g = [^{k}C_J(j) - {}^{k}C_J(j)_{g.avg}]$. Where, $TI_k(j)_{g.avg}$ is the average value (avg) of $TI_k(j)$ for a sub-set or group (g) of nodes of the same graph G ($g \in G$) that obey a given condition. We studied some collections of complex systems like MR-Nets of >40 organisms, BI-Nets of >70 ecological systems, and the SL-Net for all citations to cases of the US Supreme Court (USSC). In this case, g is not only a period (laws approved in the same year), a biological boundary (metabolic reactions in the same organism), or spatial condition (interactions in the same eco-system), but also cases citing the same USSC case. In the last problem we used a SL-Net constructed by Fowler *et al.* [218] with all cases that cite decisions of this court from 1791 to 2005. In the SL-Net of the USSC node represented a legal cases interconnected by arcs to express that the case $j^{th}$ cites the $i^{th}$ case previous to it (precedent). We constructed in total 43 sub-networks and calculated their $^{k}C_J(j)$ values and developed LNN and ANN models to predict them obtaining good results (see **Table 4**).

**Table 4.** Some QSPR models of MR-Nets, BI-Nets, and Sl-Nets

| Dataset and Model used | ANN | | Train | | | | Validation | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Li | Li = 1 | Li = 0 | % | Pr. | % | Li = 1 | Li = 0 |
| Markov-Wiener models of MR-Nets of >40 organisms | LNN 15:15-1:1  | Li = 1 | 7276 | 1985 | 78.1 | Sn | 77.9 | 21917 | 6156 |
| | | Li = 0 | 2044 | 7066 | 78.1 | Sp | 77.6 | 6227 | 21329 |
| | MLP 2:2-11-1:1  | Li = 1 | 4669 | 4559 | 50.1 | Sn | 49.7 | 13990 | 13856 |
| | | Li = 0 | 4651 | 4492 | 49.6 | Sp | 49.6 | 14154 | 13629 |
| Markov-Wiener models of BI-Nets of >70 ecosystems | MLP 13:13-13-1:1  | | Li = 1 | 4570 | 547 | 91.1 | Sn | 90.5 | 1363 |
| | | | Li = 0 | 449 | 4346 | 88.8 | Sp | 88.1 | 143 |
| | LNN 14:14-1:1  | Li = 1 | 3326 | 1710 | 66.3 | Sn | 66.1 | 995 | 603 |
| | | Li = 0 | 1693 | 3183 | 65.1 | Sp | 63.0 | 511 | 1028 |
| Markov-Wiener models of SL-Net for Spain's Financial Law system | LNN 14:14-1:1  | Li = 1 | 125 | 41 | 86.2 | Sn | 87.4 | 370 | 156 |
| | | Li = 0 | 18 | 298 | 85.4 | Sp | 87.9 | 59 | 914 |
| | MLP 14:14-14:1  | Li = 1 | 119 | 54 | 85.3 | Sn | 83.2 | 366 | 129 |
| | | Li = 0 | 24 | 285 | 87.9 | Sp | 84.1 | 63 | 941 |
| Markov-Balaban model for SL-Nets dataset of 5KCNs of USSC | MLP 18:18-10-1:1  | Li = 1 | 81225 | 51008 | 82.49 | Sn | 82.76 | 26985 | 17014 |
| | | Li = 0 | 16917 | 243415 | 82.66 | Sp | 82.7 | 5728 | 81128 |
| | LNN 18:18-1:1  | Li = 1 | 77871 | 60826 | 79.33 | Sn | 79.35 | 25950 | 20284 |
| | | Li = 0 | 20271 | 233597 | 79.33 | Sp | 79.3 | 6763 | 77858 |

## 1. CONCLUSIONS

In this work, we reviewed the recent results published about the development of MI models. We noted an evolution of MI from a simple one-target chemo-informatics algorithm for series of analogues compounds to models that are more powerful. In this sense, we illustrated the uses of the MI algorithm to solve QSPR problems in Drug-Target, Parasite-Host, Disease Spreading, Brain connectome, and Social-Legal networks. We also showed the different parameters implemented in the MI algorithm to characterize complex networks combining both classic TIs and Markov chains theory. We hope that this review may serve as inspiration to those interested on flexible, fast, and theoretically simple models for the prediction of structure-property relationships in complex systems.

## REFERENCES

1. Strogatz SH. Exploring complex networks, Nature 2001;410:268-276.
2. Estrada E. Universality in protein residue networks., Biophys J 2010;98:890-900.
3. Bonchev D. Complexity analysis of yeast proteome network, Chem. Biodivers. 2004 1:312-326.
4. Gonzalez-Diaz H. Network topological indices, drug metabolism, and distribution, Curr Drug Metab 2010;11:283-284.
5. Jeong H, Tombor B, Albert R et al. The large-scale organization of metabolic networks, Nature 2000;407:651-654.
6. Estrada E. Returnability as a criterion of disequilibrium in atmospheric reactions network, Journal of Mathematical Chemistry 2012;50:1363-1372.
7. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks, Nature 1998;393:440-442.
8. Fornito A, Zalesky A, Pantelis C et al. Schizophrenia, neuroimaging and connectomics, Neuroimage 2012;62:2296-2314.
9. Gonzalez-Diaz H, Arrasate S, Sotomayor N et al. MIANN Models in Medicinal, Physical and Organic Chemistry, Curr Top Med Chem 2013;13:619-641.
10. Christakis NA, Fowler JH. Social network sensors for early detection of contagious outbreaks, PLoS One 2010;5:e12948.
11. Hu H, Myers S, Colizza V et al. WiFi networks and malware epidemiology, Proc Natl Acad Sci U S A 2009;106:1318-1323.
12. Chen Z, Ji C. Spatial-temporal modeling of malware propagation in networks, IEEE Trans Neural Netw 2005;16:1291-1303.
13. Krause J, Ruxton GD, Krause S. Swarm intelligence in animals and humans, Trends Ecol Evol 2010;25:28-34.
14. Apicella CL, Marlowe FW, Fowler JH et al. Social networks and cooperation in hunter-gatherers, Nature 2012;481:497-501.
15. Fowler JH, Johnson TR, II JFS et al. Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court 2007.
16. Bornholdt S, Schuster HG. Handbook of Graphs and Complex Networks: From the Genome to the Internet. Wheinheim: WILEY-VCH GmbH & CO. KGa., 2003.
17. Boccaletti S, Latora V, Moreno Y et al. Complex networks: Structure and dynamics, Physics Reports 2006;424:175-308.
18. Dehmer M, Emmert-Streib F. Analysis of complex networks: from biology to linguistics. Wheinheim: Wiley-Blackwell, 2009, 462.
19. Strogatz SH. Complex systems: Romanesque networks. Nature. England, 2005, 365-366.
20. Ratti C, Sobolevsky S, Calabrese F et al. Redrawing the map of Great Britain from a network of human interactions, PLoS One 2010;5:e14248.
21. Newman ME, Watts DJ, Strogatz SH. Random graph models of social networks. Proc Natl Acad Sci U S A. United States, 2002, 2566-2572.

22. Newman M. The Structure and Function of Complex Networks, SIAM Review 2003:167-256.
23. Thomas S, Bonchev D. A survey of current software for network analysis in molecular biology, Hum Genomics 2010;4:353-360.
24. Bonchev D, Buck GA. From molecular to biological structure and back, J Chem Inf Model 2007;47:909-917.
25. Bonchev D, Rouvray DH. Complexity in Chemistry, Biology, and Ecology. New York: Springer Science+Business Media, Inc, 2005.
26. Bonchev D. Complexity analysis of yeast proteome network, Chem Biodivers 2004;1:312-326.
27. Bonchev D. On the complexity of directed biological networks, SAR and QSAR in Environmental Research 2003;14:199-214.
28. Gonzalez-Diaz H. QSAR and complex networks in pharmaceutical design, microbiology, parasitology, toxicology, cancer, and neurosciences, Current Pharmaceutical Design 2010;16:2598-2600.
29. Vina D, Uriarte E, Orallo F et al. Alignment-Free Prediction of a Drug-Target Complex Network Based on Parameters of Drug Connectivity and Protein Sequence of Receptors, Mol Pharm 2009;6:825-835.
30. 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;6:53-70.
31. Tenazinha N, Vinga S. A survey on methods for modeling and analyzing integrated biological networks, IEEE/ACM Trans Comput Biol Bioinform 2011;8:943-958.
32. Puzyn T, Leszczynski J, Cronin MTD. Recent Advances in QSAR Studies: Methods and applications. In: Leszczynski J. (ed) Challenges and advances in computational chemistry and physics. Springer, 2010, 423.
33. Estrada E, Rodríguez-Velázquez JA. Subgraph centrality in complex networks., Phys Rev E Stat Nonlin Soft Matter Phys 2005;71:056103.
34. Junker BH, Koschutzki D, Schreiber F. Exploration of biological network centralities with CentiBiN, BMC Bioinformatics 2006;7:219.
35. Gonzalez-Diaz H, Gonzalez-Diaz Y, Santana L et al. Proteomics, networks and connectivity indices, Proteomics 2008;8:750-778.
36. Kuhn M, Szklarczyk D, Franceschini A et al. STITCH 3: zooming in on protein-chemical interactions, Nucleic Acids Res 2012;40:D876-880.
37. Kuhn M, Szklarczyk D, Franceschini A et al. STITCH 2: an interaction network database for small molecules and proteins, Nucleic Acids Res 2010;38:D552-556.
38. Kuhn M, von Mering C, Campillos M et al. STITCH: interaction networks of chemicals and proteins, Nucleic Acids Res 2008;36:D684-688.
39. Zhu F, Shi Z, Qin C et al. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery, Nucleic Acids Res 2012;40:D1128-1136.
40. Hecker N, Ahmed J, von Eichborn J et al. SuperTarget goes quantitative: update on drug-target interactions, Nucleic Acids Res 2012;40:D1113-1117.
41. Gunther S, Kuhn M, Dunkel M et al. SuperTarget and Matador: resources for exploring drug-target relationships, Nucleic Acids Res 2008;36:D919-922.
42. Gaulton A, Bellis LJ, Bento AP et al. ChEMBL: a large-scale bioactivity database for drug discovery, Nucleic Acids Research 2012;40:D1100-1107.
43. Mok NY, Brenk R. Mining the ChEMBL database: an efficient chemoinformatics workflow for assembling an ion channel-focused screening library, J Chem Inf Model 2011;51:2449-2454.
44. González-Díaz H, Munteanu CR. Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks. Kerala, India: Transworld Research Network 2010, 001-212.
45. Box GEP, Jenkins GM. Time series analysis. Holden-Day, 1970.
46. Riera-Fernández P, Munteanu CR, Martín-Romalde R et al. Markov-Randic Indices for QSPR Re-Evaluation of Metabolic, Parasite-Host, Fasciolosis Spreading, Brain Cortex and Legal-Social Complex Networks, Current Bioinformatics 2013;8:401-415.

47. Riera-Fernandez P, Munteanu CR, Escobar M et al. New Markov-Shannon Entropy models to assess connectivity quality in complex networks: from molecular to cellular pathway, Parasite-Host, Neural, Industry, and Legal-Social networks, J Theor Biol 2012;293:174-188.
48. Riera-Fernandez P, Martin-Romalde R, Prado-Prado FJ et al. From QSAR models of drugs to complex networks: state-of-art review and introduction of new Markov-spectral moments indices, Curr Top Med Chem 2012;12:927-960.
49. Riera-Fernandez P, Munteanu CR, Pedreira-Souto N et al. Definition of Markov-Harary Invariants and Review of Classic Topological Indices and Databases in Biology, Parasitology, Technology, and Social-Legal Networks, Current Bioinformatics 2011;6:94-121.
50. Riera-Fernandez P, Munteanu CR, Dorado J et al. From Chemical Graphs in Computer-Aided Drug Design to General Markov-Galvez Indices of Drug-Target, Proteome, Drug-Parasitic Disease, Technological, and Social-Legal Networks, Current Computer-Aided Drug Design 2011;7:315-337.
51. Gonzalez-Diaz H, Riera-Fernandez P, Pazos A et al. The Rucker-Markov invariants of complex Bio-Systems: Applications in Parasitology and Neuroinformatics, Biosystems 2013;111:199-207.
52. Gonzalez-Diaz H, Riera-Fernandez P. New Markov-Autocorrelation Indices for Re-evaluation of Links in Chemical and Biological Complex Networks used in Metabolomics, Parasitology, Neurosciences, and Epidemiology, Journal of Chemical Information and Modeling 2012;52:3331-3340.
53. Duardo-Sanchez A, Munteanu CR, Riera-Fernández P et al. Modelling complex metabolic reactions, ecological systems, and financial-legal networks with MIANN models based on Markov-Wiener centralities, Journal of Chemical Information and Modelling 2013:submitted.
54. Patil KR, McHardy AC. Alignment-free genome tree inference by learning group-specific distance metrics, Genome Biol Evol 2013.
55. Cozzetto D, Tramontano A. Relationship between multiple sequence alignments and quality of protein comparative models, Proteins 2005;58:151-157.
56. Iliopoulos I, Tsoka S, Andrade MA et al. Evaluation of annotation strategies using an entire genome sequence, Bioinformatics 2003;19:717-726.
57. Minneci F, Piovesan D, Cozzetto D et al. FFPred 2.0: improved homology-independent prediction of gene ontology terms for eukaryotic protein sequences, PLoS One 2013;8:e63754.
58. Wood DE, Lin H, Levy-Moonshine A et al. Thousands of missed genes found in bacterial genomes and their analysis with COMBREX, Biol Direct 2012;7:37.
59. Lynch M. Intron evolution as a population-genetic process, Proc Natl Acad Sci U S A 2002;99:6118-6123.
60. Vinga S, Almeida J. Alignment-free sequence comparison-a review, Bioinformatics 2003;19:513-523.
61. Chou KC. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology, Curr Proteomics 2009;6:262-274.
62. Shen HB, Yang J, Chou KC. Methodology development for predicting subcellular localization and other attributes of proteins, Expert Rev Proteomics 2007;4:453-463.
63. Randic M, Zupan J, Balaban AT et al. Graphical representation of proteins, Chem Rev 2011;111:790-862.
64. Gonzalez-Diaz H, Dea-Ayuela MA, Perez-Montoto LG et al. QSAR for RNases and theoretic-experimental study of molecular diversity on peptide mass fingerprints of a new Leishmania infantum protein, Molecular Diversity 2010;14:349-369.
65. Aguero-Chapin G, Varona-Santos J, de la Riva GA et al. Alignment-Free Prediction of Polygalacturonases with Pseudofolding Topological Indices: Experimental Isolation from Coffea arabica and Prediction of a New Sequence, J Proteome Res 2009;8:2122-2128.
66. Gonzalez-Diaz H, Perez-Montoto LG, Duardo-Sanchez A et al. Generalized lattice graphs for 2D-visualization of biological information, Journal of Theoretical Biology 2009;261:136-147.
67. Perez-Bello A, Munteanu CR, Ubeira FM et al. Alignment-free prediction of mycobacterial DNA promoters based on pseudo-folding lattice network or star-graph topological indices, Journal of Theoretical Biology 2009;256:458-466.

68. Agueero-Chapin G, Gonzalez-Diaz H, de la Riva G et al. MMM-QSAR recognition of ribonucleases without alignment: Comparison with an HMM model and isolation from Schizosaccharomyces pombe, prediction, and experimental assay of a new sequence, Journal of Chemical Information and Modeling 2008;48:434-448.
69. Dea-Ayuela MA, Perez-Castillo Y, Meneses-Marcel A et al. HP-Lattice QSAR for dynein proteins: Experimental proteomics (2D-electrophoresis, mass spectrometry) and theoretic study of a Leishmania infantum sequence, Bioorganic & Medicinal Chemistry 2008;16:7770-7776.
70. Gonzalez-Diaz H, Agueero-Chapin G, Varona J et al. 2D-RNA-coupling numbers: A new computational chemistry approach to link secondary structure topology with biological function, Journal of Computational Chemistry 2007;28:1049-1056.
71. Gonzalez-Diaz H. Network Topological Indices, Drug Metabolism, and Distribution, Current Drug Metabolism 2010;11:283-284.
72. Gonzalez-Diaz H, Romaris F, Duardo-Sanchez A et al. Predicting Drugs and Proteins in Parasite Infections with Topological Indices of Complex Networks: Theoretical Backgrounds, Applications and Legal Issues, Current Pharmaceutical Design 2010;16:2737-2764.
73. Gonzalez-Diaz H. QSAR and Complex Networks in Pharmaceutical Design, Microbiology, Parasitology, Toxicology, Cancer and Neurosciences, Current Pharmaceutical Design 2010;16:2598-U2524.
74. González-Díaz H, Munteanu CR. Topological indices for medicinal chemistry, biology, parasitology, neurological and social networks, Kerala: Transworld Research Network 2010.
75. Hu Y, Bajorath J. Molecular scaffolds with high propensity to form multi-target activity cliffs, J Chem Inf Model 2010;50:500-510.
76. Erhan D, L'Heureux P J, Yue SY et al. Collaborative filtering on a family of biological targets, J Chem Inf Model 2006;46:626-635.
77. Namasivayam V, Hu Y, Balfer J et al. Classification of compounds with distinct or overlapping multi-target activities and diverse molecular mechanisms using emerging chemical patterns, J Chem Inf Model 2013;53:1272-1281.
78. Yildirim MA, Goh KI, Cusick ME et al. Drug-target network, Nature Biotechnology 2007;25:1119-1126.
79. Yamanishi Y, Araki M, Gutteridge A et al. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces, Bioinformatics 2008;24:i232-240.
80. Csermely P, Korcsmaros T, Kiss HJ et al. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review, Pharmacol Ther 2013;138:333-408.
81. Prado-Prado F, Garcia-Mera X, Escobar M et al. 3D MI-DRAGON: New Model for the Reconstruction of US FDA Drug-Target Network and Theoretical-Experimental Studies of Inhibitors of Rasagiline Derivatives for AChE, Current Topics in Medicinal Chemistry 2012;12:1843-1865.
82. Prado-Prado F, Garcia-Mera X, Escobar M et al. 2D MI-DRAGON: a new predictor for protein-ligands interactions and theoretic-experimental studies of US FDA drug-target network, oxoisoaporphine inhibitors for MAO-A and human parasite proteins, European Journal of Medicinal Chemistry 2011;46:5838-5851.
83. Botella-Rocamora P, Lopez-Quilez A, Martinez-Beneito MA. Spatial moving average risk smoothing, Stat Med 2013;32:2595-2612.
84. Gonzalez-Diaz H, Duardo-Sanchez A, Ubeira FM et al. Review of MARCH-INSIDE & Complex Networks Prediction of Drugs: ADMET, Anti-parasite Activity, Metabolizing Enzymes and Cardiotoxicity Proteome Biomarkers, Curr Drug Metab 2010;11:379-406.
85. Gonzalez-Diaz H, Prado-Prado F, Garcia-Mera X et al. MIND-BEST: Web Server for Drugs and Target Discovery; Design, Synthesis, and Assay of MAO-B Inhibitors and Theoretical-Experimental Study of G3PDH Protein from Trichomonas gallinae, J Proteome Res 2011;10:1698-1718.

86. Rodriguez-Soca Y, Munteanu CR, Dorado J et al. Trypano-PPI: a web server for prediction of unique targets in trypanosome proteome by using electrostatic parameters of protein-protein interactions, J Proteome Res 2010;9:1182-1190.
87. Gonzalez-Diaz H, Romaris F, Duardo-Sanchez A et al. Predicting drugs and proteins in parasite infections with topological indices of complex networks: theoretical backgrounds, applications, and legal issues, Current Pharmaceutical Design 2010;16:2737-2764.
88. Gonzalez-Diaz H, Prado-Prado FJ, Garcia-Mera X et al. MIND-BEST: web server for drugs & target discovery; design, synthesis, and assay of MAO-B inhibitors and theoretic-experimental study of G3PD protein from Trichomona gallineae, J Proteome Res 2010.
89. Munteanu CR, Vazquez JM, Dorado J et al. Complex network spectral moments for ATCUN motif DNA cleavage: first predictive study on proteins of human pathogen parasites, J Proteome Res 2009;8:5219-5228.
90. Concu R, Dea-Ayuela MA, Perez-Montoto LG et al. 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;8:4372-4382.
91. Santana L, Gonzalez-Diaz H, Quezada E et al. Quantitative structure-activity relationship and complex network approach to monoamine oxidase a and B inhibitors, Journal of Medicinal Chemistry 2008;51:6740-6751.
92. Gonzalez-Diaz H, Prado-Prado F, Ubeira FM. Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach, Curr Top Med Chem 2008;8:1676-1690.
93. Aguero-Chapin G, Gonzalez-Diaz H, de la Riva G et al. MMM-QSAR recognition of ribonucleases without alignment: comparison with an HMM model and isolation from Schizosaccharomyces pombe, prediction, and experimental assay of a new sequence, J Chem Inf Model 2008;48:434-448.
94. Aguero-Chapin G, Antunes A, Ubeira FM et al. Comparative study of topological indices of macro/supramolecular RNA complex networks, J Chem Inf Model 2008;48:2265-2277.
95. González-Díaz H, Vilar S, Santana L et al. Medicinal Chemistry and Bioinformatics – Current Trends in Drugs Discovery with Networks Topological Indices, Curr Top Med Chem 2007;7:1025-1039.
96. Ramos de Armas R, González-Díaz H, Molina R et al. Markovian Backbone Negentropies: Molecular descriptors for protein research. I. Predicting protein stability in Arc repressor mutants, Proteins 2004;56:715-723.
97. González-Díaz H, Marrero Y, Hernandez I et al. 3D-MEDNEs: an alternative "in silico" technique for chemical research in toxicology. 1. prediction of chemically induced agranulocytosis, Chem Res Toxicol 2003;16:1318-1327.
98. Gonzalez-Diaz H, Duardo-Sanchez A, Ubeira FM et al. Review of MARCH-INSIDE & complex networks prediction of drugs: ADMET, anti-parasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers, Curr Drug Metab 2010;11:379-406.
99. Batagelj V, Mrvar A. Pajek— Analysis and Visualization of Large Networks, Lecture Notes in Computer Science 2002;2265:477-478.
100. González-Díaz H, González-Díaz Y, Santana L et al. Proteomics, networks and connectivity indices, Proteomics 2008;8:750-778.
101. Santana L, Uriarte E, González-Díaz H et al. A QSAR model for in silico screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins, Journal of Medicinal Chemistry 2006;49:1149-1156.
102. Gonzalez-Diaz H, Prado-Prado F, Garcia-Mera X et al. MIND-BEST: Web Server for Drugs and Target Discovery; Design, Synthesis, and Assay of MAO-B Inhibitors and Theoretical-Experimental Study of G3PDH Protein from Trichomonas gallinae, Journal of Proteome Research 2011;10:1698-1718.
103. Gonzalez-Diaz H, Saiz-Urra L, Molina R et al. Computational chemistry approach to protein kinase recognition using 3D stochastic van der Waals spectral moments, Journal of Computational Chemistry 2007;28:1042-1048.

104. Gonzalez-Diaz H, Molina R, Uriarte E. Recognition of stable protein mutants with 3D stochastic average electrostatic potentials, Febs Letters 2005;579:4297-4301.
105. 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 6461–6473.
106. Saiz-Urra L, González-Díaz H, Uriarte E. Proteins Markovian 3D-QSAR with spherically-truncated average electrostatic potentials, Bioorganic and Medicinal Chemistry 2005;13:3641-3647.
107. González-Díaz H, Uriarte E, Ramos de Armas R. Predicting stability of Arc repressor mutants with protein stochastic moments, Bioorg Med Chem 2005;13:323-331.
108. Concu R, Podda G, Uriarte E et al. Computational chemistry study of 3D-structure-function relationships for enzymes based on Markov models for protein electrostatic, HINT, and van der Waals potentials, Journal of Computational Chemistry 2008:doi:10.1002/jcc.
109. González-Díaz H, Saiz-Urra L, Molina R et al. A Model for the Recognition of Protein Kinases Based on the Entropy of 3D van der Waals Interactions, J Proteome Res 2007;6:904-908.
110. Gonzalez-Diaz H, Saiz-Urra L, Molina R et al. Computational chemistry approach to protein kinase recognition using 3D stochastic van der Waals spectral moments, Journal of Computational Chemistry 2007;28:1042-1048.
111. Gonzalez-Diaz H, Molina R, Uriarte E. Recognition of stable protein mutants with 3D stochastic average electrostatic potentials, FEBS Letters 2005;579:4297-4301.
112. Concu R, Podda G, Uriarte E et al. Computational chemistry study of 3D-structure-function relationships for enzymes based on Markov models for protein electrostatic, HINT, and van der Waals potentials, Journal of Computational Chemistry 2009;30:1510-1520.
113. González-Díaz H, Pérez-Castillo Y, Podda G et al. Computational Chemistry Comparison of Stable/Nonstable Protein Mutants Classification Models Based on 3D and Topological Indices, J Comput Chem 2007;28:1990-1995.
114. Berca MN, Duardo-Sanchez A, González-Díaz H et al. Markov entropy for biology, parasitology, linguistic, technology, social and law networks. In: González-Díaz H., Prado-Prado F. J., García-Mera X. eds). Complex Network Entropy: From Molecules to Biology, Parasitology, Technology, Social, Legal, and Neurosciences. Kerala, India: Transworld Research Network, 2011, 127-142.
115. Riera-Fernández P, Munteanu CR, Pedreira-Souto N et al. Definition of Markov-Harary Invariants and Review of Classic Topological Indices and Databases in Biology, Parasitology, Technology, and Social-Legal Networks, Current Bioinformatics 2011;6:94-121.
116. Riera-Fernandez I, Martin-Romalde R, Prado-Prado FJ et al. From QSAR models of Drugs to Complex Networks: State-of-Art Review and Introduction of New Markov-Spectral Moments Indices, Current Topics in Medicinal Chemistry 2012.
117. StatSoft.Inc. STATISTICA (data analysis software system), version 6.0, www.statsoft.com.Statsoft, Inc., 2002.
118. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities, Proceedings of the National Academy of Sciences of the United States of America 1982;79:2554-2558.
119. Kier LB. Use of molecular negentropy to encode structure governing biological activity, Journal of Pharmaceutical Sciences 1980;69:807-810.
120. Agrawal VK, Khadikar PV. Modelling of carbonic anhydrase inhibitory activity of sulfonamides using molecular negentropy, Bioorganic & Medicinal Chemistry Letters 2003;13:447-453.
121. Graham DJ, Schacht D. Base Information Content in Organic Molecular Formulae, Journal of Chemical Information and Computer Sciences 2000;40:942.
122. Graham DJ. Information Content in Organic Molecules: Structure Considerations Based on Integer Statistics, Journal of Chemical Information and Computer Sciences 2002;42:215.
123. Graham DJ, Malarkey C, Schulmerich MV. Information Content in Organic Molecules: Quantification and Statistical Structure via Brownian Processing. , J Chem Inf Comput Sci 2004;44.

124. Graham DJ, Schulmerich MV. Information Content in Organic Molecules: Reaction Pathway Analysis via Brownian Processing, J Chem Inf Comput Sci 2004;44.
125. Graham DJ. Information Content and Organic Molecules: Aggregation States and Solvent Effects, J Chem Inf Model 2005;45.
126. Graham DJ. Information Content in Organic Molecules: Brownian Processing at Low Levels, J Chem Inf Model 2007;47:376-389.
127. Stahura FL, Godden JW, Bajorath J. Differential Shannon entropy analysis identifies molecular property descriptors that predict aqueous solubility of synthetic compounds with high accuracy in binary QSAR calculations, Journal of Chemical Information and Computer Sciences 2002;42:550-558.
128. Stahura FL, Godden JW, Xue L et al. Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binary QSAR calculations, Journal of Chemical Information and Computer Sciences 2000;40:1245-1252.
129. Roy K, Saha A. Comparative QSPR studies with molecular connectivity, molecular negentropy and TAU indicesPart I: Molecular thermochemical properties of diverse functional acyclic compounds, J Mol Model (Online) 2003;9:259-270.
130. Agrawal VK, Bano S, Khadikar PV. QSAR analysis of antibacterial activity of some 4-aminodiphenylsulfone derivatives, Acta Microbiologica et Immunologica Hungarica 2003;50:385-393.
131. Agrawal VK, Karmarkar S, Khadikar PV. QSAR study on binding affinity of PATs (rodenticides) to the [3H]-mepyramine-labelled H1 receptor in rat and guinea pig brain, SAR and QSAR in Environmental Research 2001;12:529-545.
132. Agrawal VK, Khadikar PV. QSAR studies on acylated histamine derivatives, Bioorganic and Medicinal Chemistry 2001;9:2787-2792.
133. Katritzky AR, Lomaka A, Petrukhin R et al. QSPR correlation of the melting point for pyridinium bromides, potential ionic liquids, Journal of Chemical Information and Computer Sciences 2002;42:71-74.
134. Katritzky AR, Perumal S, Petrukhin R et al. Codessa-based theoretical QSPR model for hydantoin HPLC-RT lipophilicities, Journal of Chemical Information and Computer Sciences 2001;41:569-574.
135. Strait BJ, Dewey TG. The Shannon information entropy of protein sequences, Biophysical Journal 1996;71:148-155.
136. Dima RI, Thirumalai D. Proteins associated with diseases show enhanced sequence correlation between charged residues, Bioinformatics 2004;20:2345-2354.
137. Loewenstern D, Yianilos PN. Significantly lower entropy estimates for natural DNA sequences, Journal of Computational Biology 1999;6:125-142.
138. Manke T, Demetrius L, Vingron M. Lethality and entropy of protein interaction networks, Genome Inform Ser 2005;16:159-163.
139. Mikolas P, Vyhnanek J, Skoch A et al. Analysis of fMRI time-series by entropy measures, Neuro Endocrinol Lett 2012;33:471-476.
140. Lee S, Park K, Kim D. Building a drug–target network and its applications, Expert Opin. Drug Discov. 2009;4:1-13.
141. Mestres J, Gregori-Puigjane E, Valverde S et al. The topology of drug-target interaction networks: implicit dependence on drug properties and target families, Mol Biosyst 2009;5:1051-1057.
142. Rosa da Silva M, Sun J, Ma HW et al. Metabolic networks. In: Junker B. H., Schreiber F. eds). Analysis of biological networks. New Jersey: Wiley & Sons, 2008, 233-253.
143. Lee DS, Park J, Kay KA et al. The implications of human metabolic network topology for disease comorbidity, Proceedings of the National Academy of Sciences of the United States of America 2008;105:9880-9885.
144. Huang T, Chen L, Cai YD et al. Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property, PLoS ONE 2011;6:e25297.

145. Huang T, Shi XH, Wang P et al. Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks, PLoS ONE 2010;5:e10972.
146. Baldazzi V, Ropers D, Markowicz Y et al. The carbon assimilation network in Escherichia coli is densely connected and largely sign-determined by directions of metabolic fluxes, PLoS Comput Biol 2010;6:e1000812.
147. Costa RS, Machado D, Rocha I et al. Hybrid dynamic modeling of Escherichia coli central metabolic network combining Michaelis-Menten and approximate kinetic equations, Biosystems 2010;100:150-157.
148. Gerlee P, Lizana L, Sneppen K. Pathway identification by network pruning in the metabolic network of Escherichia coli, Bioinformatics 2009;25:3282-3288.
149. Fowler ZL, Gikandi WW, Koffas MA. Increased malonyl coenzyme A biosynthesis by tuning the Escherichia coli metabolic network and its application to flavanone production, Applied and Environmental Microbiology 2009;75:5831-5839.
150. Konig R, Schramm G, Oswald M et al. Discovering functional gene expression patterns in the metabolic network of Escherichia coli with wavelets transforms, BMC Bioinformatics 2006;7:119.
151. Imielinski M, Belta C, Rubin H et al. Systematic analysis of conservation relations in Escherichia coli genome-scale metabolic network reveals novel growth media, Biophysical Journal 2006;90:2659-2672.
152. Lin H, Bennett GN, San KY. Chemostat culture characterization of Escherichia coli mutant strains metabolically engineered for aerobic succinate production: a study of the modified metabolic network based on metabolite profile, enzyme activity, and gene expression profile, Metab Eng 2005;7:337-352.
153. Ghim CM, Goh KI, Kahng B. Lethality and synthetic lethality in the genome-wide metabolic network of Escherichia coli, Journal of Theoretical Biology 2005;237:401-411.
154. Schmid JW, Mauch K, Reuss M et al. Metabolic design based on a coupled gene expression-metabolic network model of tryptophan production in Escherichia coli, Metab Eng 2004;6:364-377.
155. Light S, Kraulis P. Network analysis of metabolic enzyme evolution in Escherichia coli, BMC Bioinformatics 2004;5:15.
156. Burgard AP, Maranas CD. Probing the performance limits of the Escherichia coli metabolic network subject to gene additions or deletions, Biotechnology and Bioengineering 2001;74:364-375.
157. Edwards JS, Palsson BO. Robustness analysis of the Escherichia coli metabolic network, Biotechnology Progress 2000;16:927-939.
158. Goffeau A. The yeast genome directory, Nature 1997;387:5.
159. Burglin TR, Lobos E, Blaxter ML. Caenorhabditis elegans as a model for parasitic nematodes, International Journal for Parasitology 1998;28:395-411.
160. Consortium TCeS. Genome sequence of the nematode C. elegans: a platform for investigating biology, Science 1998;282:2012-2018.
161. Bird DM, Opperman CH. Caenorhabditis elegans: A Genetic Guide to Parasitic Nematode Biology, J Nematol 1998;30:299-308.
162. Holden-Dye L, Walker RJ. Anthelmintic drugs, WormBook 2007:1-13.
163. Muller B, Grossniklaus U. Model organisms-A historical perspective, J Proteomics 2010;73:2054-2063.
164. Poulin R. Network analysis shining light on parasite ecology and diversity, Trends Parasitol 2010;26:492-498.
165. Kotter R. Online retrieval, processing, and visualization of primate connectivity data from the CoCoMac database, Neuroinformatics 2004;2:127-144.
166. Stephan KE, Kamper L, Bozkurt A et al. Advanced database methodology for the Collation of Connectivity data on the Macaque brain (CoCoMac), Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 2001;356:1159-1186.
167. Mas-Coma S. Epidemiology of fascioliasis in human endemic areas, Journal of Helminthology 2005;79:207-216.

168. González-Díaz H, Mezo M, González-Warleta M et al. Network prediction of fasciolosis spreading in Galicia (NW Spain). In: González-Díaz H., Munteanu C. R. eds). Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks. Kerala (India): Transworld Research Network, 2010, 191-204.
169. Wasserman S, Faust K. Social network analysis: methods and applications. Cambridge: Cambridge University Press, 1999.
170. Fowler JH, Jeon S. The authority of Supreme Court precedent, Social Networks 2008;30:16-30.
171. Duardo-Sánchez A. Study of criminal law networks with Markov-probability centralities. In: González-Díaz H. (ed) Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks. Kerala, India: Bentham, 2010, 205-212.
172. Duardo-Sánchez A. Criminal law networks, markov chains, Shannon entropy and artificial neural networks. In: González-Díaz H. (ed) Complex Network Entropy: From Molecules to Biology, Parasitology, Technology, Social, Legal, and Neurosciences. Kerala, India: Bentham, 2011, 107-114.
173. Yam KL, Takhistov PT, Miltz J. Intelligent Packaging: Concepts and Applications., Journal of Food Science 2005;70:1-10.
174. Suppakul P, Miltz J, Sonneveld K et al. Active Packaging Technologies with an Emphasis on Antimicrobial Packaging and its Applications, Journal of Food Science 2003;68:408-420.
175. Rucker G, Rucker C. Walk counts, labyrinthicity, and complexity of acyclic and cyclic graphs and molecules, J Chem Inf Comput Sci 2000;40:99-106.
176. Rucker G, Rucker C. Substructure, subgraph, and walk counts as measures of the complexity of graphs and molecules, J Chem Inf Comput Sci 2001;41:1457-1462.
177. Gutman I, Rucker C, Rucker G. On walks in molecular graphs, J Chem Inf Comput Sci 2001;41:739-745.
178. Lukovits I, Trinajstic N. Atomic walk counts of negative order, J Chem Inf Comput Sci 2003;43:1110-1114.
179. Gonzalez-Diaz H, Riera-Fernandez P, Pazos A et al. The Rucker-Markov invariants of complex Bio-Systems: Applications in Parasitology and Neuroinformatics, Biosystems 2013.
180. Moreau G, Broto P. Autocorrelation of a topological structure: A new molecular descriptor, Nouv J Chim 1980;4:359-360.
181. Moreau G, Broto P. Autocorrelation of molecular structures, application to SAR studies, Nouv J Chim 1980;4:757–764.
182. Broto P, Moreau G, Vandycke C. Molecular structures: perception, autocorrelation descriptor and sar studies: system of atomic contributions for the calculation of the n-octanol/water partition coefficients, European Journal of Medicinal Chemistry 1984;19:71-78.
183. Moro S, Bacilieri M, Cacciari B et al. Autocorrelation of molecular electrostatic potential surface properties combined with partial least squares analysis as new strategy for the prediction of the activity of human A(3) adenosine receptor antagonists, Journal of Medicinal Chemistry 2005;48:5698-5704.
184. Fernández L, Caballero J, Abreu JI et al. 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.
185. Caballero J, Garriga M, Fernandez M. 2D Autocorrelation modeling of the negative inotropic activity of calcium entry blockers using Bayesian-regularized genetic neural networks, Bioorganic and Medicinal Chemistry 2006;14:3330-3340.
186. Caballero J, Fernández L, Garriga M et al. Proteometric study of ghrelin receptor function variations upon mutations using amino acid sequence autocorrelation vectors and genetic algorithm-based least square support vector machines, Journal of Molecular Graphics and Modelling 2007;26:166-178.
187. Caballero J, Fernandez L, Abreu JI et al. Amino Acid Sequence Autocorrelation vectors and ensembles of Bayesian-Regularized Genetic Neural Networks for prediction of conformational stability of human lysozyme mutants, J Chem Inf Model 2006;46:1255-1268.

188. Han P, Zhang X, Norton RS et al. Large-scale prediction of long disordered regions in proteins using random forests, BMC Bioinformatics 2009;10:8.
189. Li ZR, Lin HH, Han LY et al. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence, Nucleic Acids Res 2006;34:W32-37.
190. Munteanu CR, Gonzalez-Diaz H, Borges F et al. Natural/random protein classification models based on star network topological indices, Journal of Theoretical Biology 2008;254:775-783.
191. Modha DS, Singh R. Network architecture of the long-distance pathways in the macaque brain, Proceedings of the National Academy of Sciences of the United States of America 2010;107:13485-13490.
192. Wiener H. Structural determination of paraffin boiling points, Journal of the American Chemical Society 1947;69:17-20.
193. Wiener H. Correlation of heats of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons, Journal of the American Chemical Society 1947;69:2636–2638.
194. Wiener H. Relation of the physical properties of the isomeric alkanes to molecular structure. Surface tension, specific dispersion, and critical solution temperature in aniline, The Journal of Physical and Colloid Chemistry 1948;52:1082-1089.
195. Wiener H. Vapor pressure–temperature relationships among the branched paraffin hydrocarbons, The Journal of Physical and Colloid Chemistry 1948;52:425-430.
196. Hosoya H. Topological index,a newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons, Bulletin of the Chemical Society of Japan 1971;44:2332-2339.
197. Hosoya H. Mathematical meaning and importance of the topological index Z, Croatica Chemica Acta 2007;80:239-249.
198. Consonni V, Todeschini R. Molecular descriptors. In: Puzyn T., Leszczynski J., Cronin M. T. D. eds). Recent advances in QSAR studies: Methods and applications. Springer, 2010, 29-102.
199. Harary F. Status and contrastatus, Sociometry 1959;22:23-43.
200. Diudea MV, Gutman I. Wiener-type topological indices, Croatica Chemica Acta 1998;71:21-51.
201. Rosenblatt F. Principles of neurodynamics; perceptrons and the theory of brain mechanisms. Washington: Spartan Books, 1962.
202. T BA. HIGHLY DISCRIMINATING DISTANCE-BASED TOPOLOGICAL INDEX, Chem. Phys. Lett. 1981;89:399-404.
203. Sharma SK, Kumar P, Narasimhan B et al. Synthesis, antimicrobial, anticancer evaluation and QSAR studies of, Eur J Med Chem 2012;48:16-25.
204. Thakur A, Thakur M, Khadikar PV et al. QSAR study on benzenesulphonamide carbonic anhydrase inhibitors: topological, Bioorg Med Chem 2004;12:789-793.
205. Krawczuk A, Voelkel A, Lulek J et al. Use of topological indices of polychlorinated biphenyls in structure-retention, J Chromatogr A 2003;1018:63-71.
206. Yadav S, Kumar P, De Clercq E et al. 4-[1-(Substituted aryl/alkyl carbonyl)-benzoimidazol-2-yl]-benzenesulfonic acids, Eur J Med Chem 2010;45:5985-5997.
207. Naik PK, Dubey A, Kumar R. Development of predictive quantitative structure-activity relationship models of, J Biomol Screen 2010;15:1194-1203.
208. Fernandes JP, Pasqualoto KF, Felli VM et al. QSAR modeling of a set of pyrazinoate esters as antituberculosis prodrugs, Arch Pharm (Weinheim) 2010;343:91-97.
209. Panaye A, Doucet JP, Devillers J et al. Decision trees versus support vector machine for classification of androgen, SAR QSAR Environ Res 2008;19:129-151.
210. Ma XL, Chen C, Yang J. Predictive model of blood-brain barrier penetration of organic compounds, Acta Pharmacol Sin 2005;26:500-512.
211. Dashtbozorgi Z, Golmohammadi H. Quantitative structure-property relationship modeling of water-to-wet butyl, J Sep Sci 2010;33:3800-3810.

212. Jalali-Heravi M, Fatemi MH. Prediction of thermal conductivity detection response factors using an artificial neural network, J Chromatogr A 2000;897:227-235.
213. Dehmer M, Grabner M, Varmuza K. Information indices with high discriminative power for graphs, PLoS One 2012;7:e31214.
214. Basak SC, Mills D, Gute BD et al. Use of mathematical structural invariants in analyzing combinatorial libraries: a, Curr Comput Aided Drug Des 2010;6:240-251.
215. Ratkiewicz A, Truong TN. Application of chemical graph theory for automated mechanism generation, J Chem Inf Comput Sci 2003;43:36-44.
216. Randic M, Pompe M. The variable molecular descriptors based on distance related matrices, J Chem Inf Comput Sci 2001;41:575-581.
217. Duardo-Sanchez A, Munteanu CR, Pazos A et al. ANN Modeling of Complex Networks of Biochemical Reactions, Ecosystems, and U.S. Supreme Court Citations with New Balaban-Markov Centralities, Complexity 2013:submitted.
218. Fowler JH, Jeon S. The authority of Supreme Court precedent, Social Networks 2008; 30:16-30.