Métodos grafo-numéricos de relevancia para la bioinformática. Aplicaciones en la biotecnología vegetal y en el descubrimiento de fármacos.

Autoría principal
Guillermin Agüero Chapin[1].

Otros autores
Humberto González Díaz[2], Reinaldo Molina Ruiz[1], Aminael Sánchez Rodríguez[3], Gisselle Pérez Machado[1], Yunierkis Pérez Castillo[1], Aliuska Morales Helguera[1], Agostinho Antunes[4], Vitor Vasconselos[4].

Colaboradores
Javier Varona Santos, Eugenio Uriarte, Yennny González Díaz, Giovanna Delogu, Lourdes Santana, Gianni Podda, Gustavo de la Riva, Roberto I. Vazquez-Padrón, Florencio M. Ubeira, Kuo-Chen Chou, Tomás González-Villa, María A. Dea-Ayuela, Lázaro G. Pérez-Montoto, Francisco J. Prado-Prado, Francisco Bolas-Fernández, Pedro I. Hidalgo-Yanes, Kathleen Marchal, Emanuel Maldonado, Christian Munteanu, Aliuska Duardo, Grace Patlewicz, Alberto López-Díaz.

Entidad ejecutora principal
[1]Centro de Bioactivos Químicos, Universidad Central "Marta Abreu" de Las Villas, Carretera a Camajuaní Km 5 ½, Santa Clara, 54830, Cuba.

Entidades participantes
[2]Departmento de Química Orgánica, Universidad Heriko Euskal, Leioa, 48940, Biskaia, País Vasco, España.
[3] Departamento de Ciencias Naturales, Universidad Técnica Particular de Loja, San Cayetano Alto, S/N, Loja, Ecuador.
[4]CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Rua dos Bragas, 177, 4050-123 Porto, Portugal.

Autor para correspondencia
Guillermin Agüero Chapin.
Centro de Bioactivos Químicos, Universidad Central "Marta Abreu" de Las Villas.
Carretera a Camajuaní Km 5 ½, Santa Clara, 54830, Cuba.
email: chapin@uclv.edu.cu

Aporte científico de cada autor al resultado

- ✓ **Guillermin Agüero Chapin** (30%): Diseñó y participó en el 100% de la investigación y además en la escritura de todos los artículos publicados.
- ✓ **Humberto González Díaz** (20%): Diseñó y participó en el 60% de la investigación y además en la escritura de varios artículos publicados.
- ✓ **Reinaldo Molina Ruiz** (20%): Programó los software TI2BioP y MARCH-INSIDE.
- ✓ **Aminael Sánchez Rodríguez** (5%): Condujo los análisis basados en alineamientos y participó en la escritura de varios artículos.
- ✓ **Gisselle Pérez Machado** (5%): Llevó a cabo la validación experimental de los modelos.
- ✓ **Aliuska Morales Helguera** (5%): Participó en los análisis estadísticos.

- ✓ **Yunierkis Pérz Castillo** (5%): Gestionó y editó las bases de secuencias.
- ✓ **Agostinho Antunes** (5%): Analizó los resultados y participó en la escritura de los artículos.
- ✓ **Vitor Vasconselos** (5%): Analizó los resultados y participó en la escritura de los artículos.

Contribución de la entidad cubana (Centro de Bioactivos Químicos)

El Centro de Bioactivos Químicos aportó su experiencia en la química computacional para el desarrollo y aplicación de ambos programas (**TI2BioP y MARCH-INSIDE). TI2BioP** es el primer recurso grafo-numérico disponible libremente en internet, alternativo a los métodos clásicos para hacer búsquedas de genes y proteínas a baja diversidad estructural y además con potencial para inferir relaciones evolutivas entre familias de genes/proteínas altamente divergentes. Su desarrollo estuvo totalmente inspirado en los resultados previos alcanzados con la metodología **MARCH-INSIDE** dentro de la genómica y proteómica funcional en plantas. La aplicación de **TI2BioP** está dirigida a la búsqueda de genes/proteínas que intervienen directa o indirectamente en el descubrimiento de nuevos compuestos bioactivos.

Contribución de la entidad extanjera (CIIMAR)

El Centro Interdisciplinario de Investigación Marina y Ambiental de la Universidade do Porto aportó su experiencia experimental en estudios genómicos y proteómicos y en las técnicas aplicadas a la evolución molecular. Dicha entidad financió las publicaciones de libre acceso así como los insumos para llevar a cabo la parte experimental de validación del método.

Resumen

En este trabajo desarrollamos un nuevo método grafo-numérico llamado **TI2BioP** (**T**opological **I**ndices **to BioP**olymers) como herramienta bioinformática para la clasificación funcional e inferencias filogenéticas en familias de genes/proteínas de relevancia en el descubrimiento de fármacos. Dicha investigación está inspirada en la metodología **MARCH-INSIDE** (**Mar**kov **Ch**ain **I**nvariants for **N**etwork **S**election **& De**sign), la cual fue desarrollada y aplicada por nuestro grupo a la identificación de genes y proteínas de importancia para la Biotecnología Vegetal. Ambos enfoques grafo-teóricos han sido importantes en la bioinformática por su sensibilidad en la identificación de genes y proteínas difíciles de detectar por los métodos actuales de búsqueda de secuencias. Ambas herramientas computacionales fueron desarrolladas en el Centro de Bioactivos Químicos y como resultado de su aplicación se produjeron 13 publicaciones internacionales con más de 20 colaboradores nacionales e internacionales. Se publicó además un libro y dos capítulos publicados por editoriales de EUA e India. Los resultados de la investigación fueron presentados en 10 eventos internacionales y además motivo de tesis de grado, maestría y doctorado; además de un premio ACC Provincial en el 2013. El nuevo software **TI2BioP** está en proceso de registro pero ya está de libre acceso con fines de investigativos en el sitio http://ti2biop.sourceforge.net/.

Comunicación Corta

Motivación: Existen varias herramientas bioinformáticas para buscar señales funcionales determinantes en familias de genes y proteínas, pero todas ellas generalmente se basan en alguna medida de similitud de secuencia[1]. Los métodos basados en similitud tienen como núcleo algorithmos de alineamiento de secuencias que comúnmente se aplican para hacer búsquedas por homología, clasificar estructural

y funcionalmente una secuencia problema e inferir relaciones evolutivas[2]. A pesar que dichos métodos presentan una interface amistosa con el usuario y han evolucionado hacia algoritmos más precisos, aún muestran un pobre desempeño en detectar miembros funcionales en familias de genes y proteínas altamente diversas[3,4]. Por otra parte las inferencias filogenéticas que dependen de alineamientos múltiples de secuencias no son confiables cuando los genes/proteínas muestran una similitud funcional pero han divergido enormemente[2,4]. Como consecuencia de este hecho, se han desarrollado varios métodos independientes de alineamientos para superar dicho problema. La generalidad de ellos se han enfocado en explotar la composición de nucleótidos y aminoácidos y así obtener clasificadores libres de alineamiento que son usados por métodos de aprendizaje automatizado para desarrollar algoritmos de clasificación[5,6]. Actualmente la teoría de grafos esta siendo extendida a la bioinformática a través de la introducción de grafos (2D) bi-dimensionales para análisis comparativos de ADN/ARN y proteínas sin necesidad de recurrir a los métodos tradicionales basados en similitud de secuencias[7]. A pesar que varios métodos grafo-numéricos han sido desarrollados para abordar este tipo de análisis[8]; muy pocos han sido aplicados con un objetivo práctico para realiar análisis masivos de secuencias. Especialmente en conjunto de datos donde los métodos tradicionales muestran bajo desempeño durante la predicción funcional y análisis filogéneticos[9-11]. Por consiguiente desarrollamos una nueva herramienta denominada **TI2BioP** (**T**opological **I**ndices **to BioP**olymers)[12], para enfrentar limitaciones de los métodos tradicionales basados en alineamientos. **TI2BioP** está inspirado en experiencias anteriormente alcanzadas por la metodología **MARCH-INSIDE** (**Mar**kov **Ch**ain **I**nvariants for **N**etwork **S**election **& De**sign)[13] en la anotación funcional de genes y proteínas de importancia para la biotecnología vegetal[9-11, 14-17]**.** La metodología **MARCH-INSIDE** fue el primer recurso grafo-numérico, libre de alineamiento, desarrollado por nuestro grupo para establecer relaciones de estructura-función biológica en genes y proteínas de origen vegetal. A pesar de sus resultados alentadores en la detección de nuevos miembros en familias relacionadas con la maduración de frutos tropicales como la 1-aminociclopropano-1-ácido carboxílico (ACC) oxidasas y las polygalacturonasas y en otras como los genes ribosomales 18s; existía la necesidad de desarrollar una nueva herramienta más simple y de acceso público para hacer frente a problemas angulares de la bioinformática actual. De esta forma, **TI2BioP** extiende el cálculo de índices topológicos (ITs) simples introducidos por Estrada *y col*. para moléculas orgánicas pequeñas[18] a biopolímeros como ADN, ARN y proteínas representados a partir de diferentes enfoques gráficos implementados en el **MARCH-INSIDE**[19]. Dichos ITs fueron evaluados en predecir funciones biológicas de familias de genes y proteínas altamente diversas y para inferir relaciones filogenéticas a baja simlitud de secuencia. Se escogieron familias que tenían en común la baja similitud entre sus miembros y cierta relevancia para el descubrimiento de nuevos fármacos, ya sea por proveer metabolitos secundarios con actividad biológica (bacteriocinas y péptidos no ribosomales) o por representar blancos para el diseño de fármacos (RNasa III).

Resultados: Desarrollamos el sofware **TI2BioP** que calcula los momentos espectrales como simples ITs a partir de diferentes enfoques gráficos que caracterizan ADN, ARN y proteínas. **TI2BioP** como se mencionó anteriormente es el resultado de

nuestras experiencias anteriores con la metodología **MARCH-INSIDE** en el campo de la genómica y proteómica en plantas**.** A pesar que los ITs calculados por ambos recursos grafo-teóricos son diferentes por definición fueron desarrollados y aplicados para propósitos similares por lo que existe una línea de investigación coherente desde el año 2006 hasta la actualidad. **TI2BioP** utiliza la representación Cartesiana para genes y proteínas, la representación termodinámica del plegamiento de ADN y ARN, anteriormente implementadas en la plataforma **MARCH-INSIDE,** para el cálculo de los ITs. También se implementó de forma novedosa los mapas de cuatro colores para ADN, ARN y proteínas con la misma finalidad. De dichas representaciones pudimos derivar los momentos espectrales para su aplicación en la bioinformática y así enfrentar el bajo desempeño que muestran los métodos actuales para lidear con familias de genes y proteínas de elevada diversidad. **TI2BioP** es el primer recurso grafo-numérico libremente disponible en http://ti2biop.sourceforge.net/, que aborda esta temática y está contando ya con su version 2.0, la cual ha sido utiizada por más de 15 usuarios a nivel internacional[12].

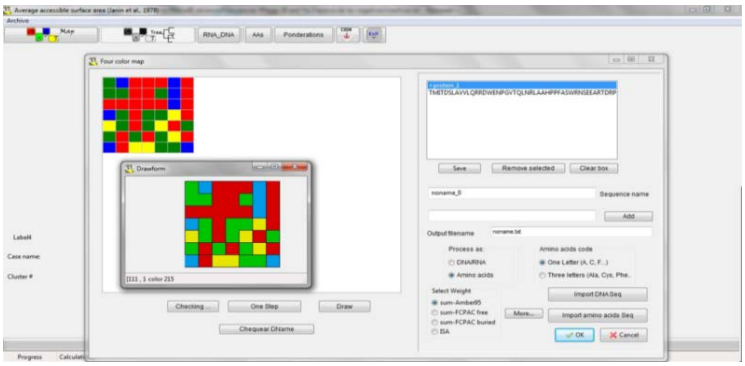

Figura 1. Vista de **TI2BioP** para la representación de los mapas de cuatro colores para proteínas.

Es válido mencionar que **MARCH-INSIDE** fue el pionero, pero con un enfoque mas quimio que bioinformático puesto que no fue concebido para apoyar o resolver problemas directos de la bioinformática y sí del diseño de fármacos. Sin embargo dicha metodología fue introducida para anotar funciones de genes y proteínas de plantas sin acudir a los algoritmos de alineamientos usando enfoques gráficos ya reportados y algunos introducidos por nosotros. De esta forma pudimos anotar nuevos miembros para clases involucradas en la maduración de frutos, que luego serían utilizados por la ingeniería genética para el control de la maduración de frutos. Los nuevos miembros predichos, aislados y registrados en bancos de genes/proteínas fueron las ACC oxidasas y poligalacturonasas de variedades de café[10,15] y guayaba[14,20] asi como ribonucleases III de la levadura *Schizosaccharomyces pombe*[9] y del patógeno *Leishmania infantum*[17].

Por otra parte**,TI2BioP** fue evaluado en la clasificación funcional de las bacteriocinas proteicas[21], RNasas III[22], ITS2[23] y los dominios de Adenilación de las Sintetasas de Péptidos No Ribosomales (SPNRs) . Los ITs generados por **TI2BioP** son variables de entrada para técnicas estadísticas tales como Discriminante Lineal (DL), Redes de Inteligencia Artificial (RIA) y Arboles de Decisión (AD), que permiten el desarrollo de

modelos de clasificación altamente predictivos para enfrentar la diversidad de secuencias en las clases. La utilidad de dichos modelos fue demostrada en detectar nuevos miembros pertenecientes a cada clase involucrada en el estudio. Las predicciones estuvieron apoyadas por evidencias experimentales y por predicciones realizadas por métodos avanzados de alineamientos. **TI2BioP** detectó un dominio tipo bacteriocina con posible uso en la industria biotecnológica, el cual no había sido reconocido por los algoritmos bioinformáticos actuales, solamente por la experimentación[24]. Dicha metodología contribuyó además a la detección de nuevas variantes de dominios de Adenilación en el proteoma de la cyanobacteria *Microcystis aeruginosa,* como forma de detectar nuevos grupos de SPNRs y por consiguiente participa en el descubrimiento de rutas biosintéticas de nuevos péptidos con actividad biológica. El descubrimiento de nuevos candidatos naturales a fármacos a nivel genómico y proteómico es de relevancia social debido a la necesidad de encontrar nuevos antibióticos que sean efectivos contra cepas resistentes y antitumorales con menos efectos adversos. Desde el punto de vista económico, la búsqueda "in silico" brinda un considerable ahorro de recursos en comparación con el tamizaje químico-microbiológico que tradicionalmente se aplica para el hallazgo de fármacos en fuentes naturales.

La tabla 1 muestra los mejores modelos libres de alineamientos obtenidos para la clasificación funcional de las familias objeto de estudio y el procedimiento llevado a cabo para detectar o anotar funcionalmente nuevos miembros.

Tabla 1. Resumen de la construcción y aplicación de los modelos obtenidos por **MARCH-INSIDE y TI2BioP** para la clasificación funcional de cada familia de genes/proteínas estudiadas y la detección de nuevos miembros.

Clase gen/proteína	Método grafo-numérico	Tipo de Grafo 2D	Modelo	Nuevo(a) gene/proteína	Procedimento de Anotación
ACC oxidasa	**MARCH-INSIDE**	Termodinámica	DL	ACC oxidasa de *Coffea arabica*	Predicción libre de alineamiento y Predicción basada en similitud
Poligalacturonasa	**MARCH-INSIDE**	Cartesiana	DL	Poligalacturonasa de *Psidium guajava L*	Predicción libre de alineamiento y Evidencias experimentales
Poligalacturonasa	**MARCH-INSIDE**	Cartesiana	DL	Poligalacturonasa de *Coffea arabica*	Predicción libre de alineamiento y Evidencias experimentales
Ribonucleasa III (RNasa III)	**MARCH-INSIDE**	Cartesiana	DL	RNase III de *Schizosaccharomyces pombe*	Predicción libre de alineamiento , basada en homologia y Evidencias experimentales

Bacteriocinas Proteicas	**TI2BioP**	Cartesiana	DL	Dominio C-terminal Cry 1Ab *Bacillus thuringiensis*	Predicción libre de alineamiento y Evidencias experimentales
ITS2 Genómico	**TI2BioP**	Cartesiana y Termodinámica	RIA	ITS2 genómica *Petrakia* sp.	Predicción libre de alineamiento y basada en homología
RNasa III	**TI2BioP**	Cartesiana	AD	RNasa III *E coli BL21 subcepa CG 1208*	Predicción libre de alineamiento y Evidencias experimentales
Dominios Adenilación de SPNRs	**TI2BioP**	Mapa de cuatro colores	AD	5 coincidencias en el proteoma de *Microcystis aeruginosa*	Pendiente a registro

El desempeño de los modelos libres de alineamientos fue generalmente comparado con procedimientos avanzados de alineamientos tales como InterPro y perfiles HMM en la detección de miembros de las clases seleccionadas (Tabla 2).

Tabla 2. Comparación de ambos métodos a través del desempeño en la clasificación del subconjunto de prueba (Sensibilidad). Identificación de nuevos miembros por ambos tipos de modelos. Los algoritmos de alinemiento alcanzan la máxima sensibilidad solo cuando se aplican estrategias complejas.

	Métodos libres de Alineamientos			**Procedimientos de Alineamientos**		
Clase Gen/proteína	**Método grafo-numérico**	**Sensibilidad subc. prueba**	**Detección Nuevos Miembros**	**Algoritmo Alineamiento**	**Sensibilidad subc. prueba**	**Detección Nuevos Miembros**
RNasa III	**MARCH-INSIDE**	100%	Con Significación	Perfil HMM	98.75%	Con Significación
Celulasa	**MARCH-INSIDE**	100%	--	Perfil HMM	82.99%	--
Bacteriocinas Proteicas	**TI2BioP**	66.67%	Con Significación	InterPro	60.2%	No Detección
ITS2 Genómica	**TI2BioP**	92.59%	Con Significación	Perfil HMM (MAFFT)	66.66%	Con Significación
RNasa III	**TI2BioP**	96.07%	Con Significación	Perfil HMM (modificado)	100%	Con Significación
Dominios Adenilación	**TI2BioP**	100%	Con Significación	Perfil HMM	100%	Con Significación

El enfoque gráfico de **TI2BioP** fue también usado para visualizar relaciones funcionales ocultas con miembros distantes como fue el caso de las bacteriocina y el dominio C-terminal de la endotoxina Cry1AB del *Bacillus thuringiensis*[21]. La

metodología además pemitió llevar a cabo una taxonomía molecular conducida por los ITs. La clase ITS2 fue usada para complementar la ubicación taxonómica del hongo del género *Petrakia* sp. (posible productor de compuestos bioactivos) con técnicas libres de alineamietntos aplicadas a las inferencias filogéneticas. El controversial hongo fue ubicado por primera vez dentro del subfilo *Pezizomycotina* y la clase *Dothideomycetes*[23]*.* La metodología superó también a los métodos tradicionales en detectar dominios de Adenilación de las SPNRs, solamente los algoritmos complejos de alineamientos (perfiles HMM) fueron capaces de tener igual desempeño. **TI2BioP** ensamblado con dichos algoritmos fue capaz de detectar posibles nuevas formas de dominios de Adenilación en el proteoma de la cianobacteria *Microcystis aeruginosa* como estrategia para descubir nuevas vías de producción de oligopéptidos con interés farmacéutico[25].

Referencias

[1] Holm, L. and C. Sander, Protein folds and families: sequence and structure alignments. Nucleic Acids Res, 1999. 27(1): p. 244-7.

[2] Hohl, M. and M.A. Ragan, Is multiple-sequence alignment required for accurate inference of phylogeny? Syst Biol, 2007. 56(2): p. 206-21.

[3] Dobson, P.D. and A.J. Doig, Distinguishing Enzyme Structures from Non-enzymes Without Alignments. J. Mol. Biol., 2003. 330: p. 771–783.

[4] Schwarz, R.F., et al., Evolutionary Distances in the Twilight Zone—A Rational Kernel Approach. PLoS ONE, 2010. 5(12).

[5] Strope, P.K. and E.N. Moriyama, Simple alignment-free methods for protein classification: a case study from G-protein-coupled receptors. Genomics, 2007. 89(5): p. 602-12.

[6] Chou, K.C., Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins, 2001. 43(3): p. 246-55.

[7] Randic, M., et al., Graphical representation of proteins. Chem Rev, 2011. 111(2): p. 790-862.

[8] Gonzalez-Diaz, H., et al., Generalized lattice graphs for 2D-visualization of biological information. J Theor Biol, 2009. 261(1): p. 136-47.

[9] Aguero-Chapin, 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(2): p. 434-48.

[10] Aguero-Chapin, G., 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(4): p. 2122-8.

[11] Agüero-Chapin, G., et al., Network entropies classification of fungi and bacteria cellulases of interest for biotechnology, in Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks, H.G.-D.a.C.R. Munteanu, Editor 2010, Transworld Research Network: Kerala.

[12] Molina, R., G. Agüero-Chapin, and M.P. Pérez-González, TI2BioP (Topological Indices to BioPolymers) version 2.0.2011: Molecular Simulation and Drug Design (MSDD), Chemical Bioactives Center, Central University of Las Villas, Cuba.

[13] González-Díaz H, Molina-Ruiz R, and Hernandez I, MARCH-INSIDE v3.0 (MARkov CHains INvariants for SImulation & DEsign), 2007. p. Windows supported version under request to the main author contact email: gonzalezdiazh@yahoo.es.

[14] Aguero-Chapin, G., et al., Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from Psidium guajava L. FEBS Lett, 2006. 580(3): p. 723-30.

[15] Gonzalez-Diaz, H., et al., 2D-RNA-coupling numbers: a new computational chemistry approach to link secondary structure topology with biological function. J Comput Chem, 2007. 28(6): p. 1049-56.

[16] Aguero-Chapin, G., et al., Comparative study of topological indices of macro/supramolecular RNA complex networks. J Chem Inf Model, 2008. 48(11): p. 2265-77.

[17] Gonzalez-Diaz, H., et al., QSAR for RNases and theoretic-experimental study of molecular diversity on peptide mass fingerprints of a new Leishmania infantum protein. Mol Divers, 2010. 14(2): p. 349-69.

[18] Estrada, E., On the topological sub-structural molecular design (TOSS-MODE) in QSPR/QSAR and drug design research. SAR QSAR Environ Res, 2000. 11(1): p. 55-73.

[19] Estrada, E., Spectral Moments of the Edge-Adjacency Matrix of Molecular Graphs. 2. Molecules Containing Heteroatoms and QSAR Applications. J Chem Inf Comput Sci, 1997. 37: p. 320-328.

[20] Gonzalez-Diaz, H., et al., 2D RNA-QSAR: assigning ACC oxidase family membership with stochastic molecular descriptors; isolation and prediction of a sequence from Psidium guajava L. Bioorg Med Chem Lett, 2005. 15(11): p. 2932-7.

[21] Aguero-Chapin, G., et al., TI2BioP: Topological Indices to BioPolymers. Its practical use to unravel cryptic bacteriocin-like domains. Amino Acids, 2011. 40(2): p. 431-42.

[22] Aguero-Chapin, G., et al., Non-linear models based on simple topological indices to identify RNase III protein members. J Theor Biol, 2011. 273(1): p. 167-178.

[23] Aguero-Chapin, G., et al., An alignment-free approach for eukaryotic ITS2 annotation and phylogenetic inference. PLoS One, 2011. 6(10): p. e26638.

[24] Vazquez-Padron, R.I., et al., Cryptic endotoxic nature of Bacillus thuringiensis Cry1Ab insecticidal crystal protein. FEBS Lett, 2004. 570(1-3): p. 30-6.

[25] Aguero-Chapin, G., et al., Exploring the Adenylation Domain Repertoire of Nonribosomal Peptide Synthetases Using an Ensemble of Sequence-Search Methods. PLoS One, 2013. 8(7).

[26] Estrada, E., Spectral Moments of the Edge Adjacency Matrix in Molecular Graphs. 1. Definition and Applications to the Prediction of Physical Properties of Alkanes. J Chem Inf Comput Sci, 1996. 36: p. 844-849.

Publicaciones

- ✓ **Agüero-Chapin, G**; González-Díaz, H; Molina R; Varona-Santos, J; Uriarte, E; González-Díaz, Y. Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from *Psidium guajava* L. ***FEBS Letters* 580 (2006) 723–730**
- ✓ Gonzalez-Diaz H, **Agüero-Chapin G**, Varona J, Molina R, Delogu G, Santana L, Uriarte E, Podda G. 2D-RNA-coupling numbers: A new computational chemistry approach to link secondary structure topology with biological function. ***J Comput Chem*. 2007 Apr 30;28(6):1049-56**
- ✓ **Agüero-Chapin G**, González-Díaz H, de la Riva G, Rodríguez E, Sánchez-Rodríguez A, Podda G, Vázquez-Padrón RI. 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(2) 434-448.**
- ✓ **Agüero-Chapin G**, Antunes A, Ubeira F; Chou, KC; González-Díaz H. Comparative Study of Topological Indices of Macro/Supra-molecular RNA Complex Networks. ***J. Chem. Inf. Model.* 2008 48(11)**: **2265-77**
- ✓ **Agüero-Chapin G**, Varona-Santos J, de la Riva GA, Antunes A, González-Villa T, Uriarte E, González-Díaz H. Alignment-Free Prediction of Polygalacturonases with Pseudofolding Topological Indices: Experimental Isolation from *Coffea arabica* and Prediction of a New Sequence. ***Journal of Proteome Research.* 2009 8(4):2122-2128**
- ✓ González-Díaz H, **Agüero-Chapin G**, Munteanu C.R, Prado-Prado F, Chou KC, Duardo-Sanchez A, Patlewicz G, and López-Díaz A: *Alignment-free models in Plant Genomics: Theoretical, Experimental, and Legal issues* ISBN: 978-1-61668-333-7, Retail Series: Agriculture Issues and Policies, **Pub. Date**: 2010. https://www.novapublishers.com/catalog/product_info.php?products_id=12947
- ✓ González-Díaz H, **Agüero-Chapin G**, Munteanu C.R, Prado-Prado F, Chou KC, Duardo-Sanchez A, Patlewicz G, and López-Díaz A (2011) In: Advances in Genetics Research (vol. 1), Ed. Maria A. Osborne, **Nova Science Publishers**, Inc., NY, USA, ISBN: 978-1-60692-638-3. Chapter 2: Alignment-free models in Plant Genomics: Theoretical, Experimental, and Legal issues. Web: https://www.novapublishers.com/catalog/product_info.php?products_id=9340
- ✓ **Agüero-Chapin, G,** Sánchez-Rodríguez, A, Antunes, A, de la Riva G.A and González Díaz H. (2010) In: Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks. Editors: H. González-Díaz and C.R. Munteanu. **Transworld Research Network**, Kerala, India. ISBN 978-81-7895-989-9. Chapter 5: Network entropies classification of fungi and bacteria of fungi and bacteria cellulases of interes for biotechnology. page: 69-95.
- ✓ González-Díaz H, Dea-Ayuela MA, Pérez-Montoto LG, Prado-Prado FJ, **Agüero-Chapín G**, Bolas-Fernández F, Vazquez-Padrón RI, Ubeira FM. 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.**
- ✓ **Agüero-Chapin G**, Pérez-Machado G, Molina-Ruiz R, Pérez-Castillo Y, Morales-Helguera A, Vasconcelos V and Antunes A. TI2BioP: Topological Indices to BioPolymers. Its practical use to unravel cryptic bacteriocin-like domains. ***Amino Acids.* 2011; 40(2):431-42.**

- ✓ **Agüero-Chapin G,** de la Riva GA, Molina-Ruiz R, Sánchez-Rodríguez A, Pérez-Machado G, Vasconcelos V and Antunes A. Non-linear models based on simple topological indices to identify RNase III protein members. ***Journal of Theoretical Biology***. **2011; 273(1):167-78**.
- ✓ **Agüero-Chapin G,** Sánchez-Rodríguez A, Hidalgo-Yanes PI, Pérez-Castillo Y, Molina-Ruiz R, Marchal K, Vasconcelos V and Antunes A. An alignment-free approach for eukaryotic ITS2 annotation and phylogenetic inference. ***PLoS ONE*** **2011;6(10).**
- ✓ **Agüero-Chapin G,** Molina-Ruiz R, Maldonado E, de la Riva GA, Vasconcelos V and Antunes A. Exploring the adenylation domain repertoire of nonribosomal peptide synthetases using an ensemble of sequence-search methods. ***PLoS ONE*** **2013; 8(7).**