Tesis Doctoral

Modelado y Explotación De Redes Para La Caracterización y Predicción En Sistemas Biomédicos

Aníbal Bueno Amorós

UNIVERSIDAD DE MÁLAGA

Director: Dr. Juan Antonio García Ranea
Facultad de Ciencias
Fundamentos celulares y moleculares de los seres vivos

UNIVERSIDAD
DE MÁLAGA

AUTOR: Aníbal Bueno Amorós

http://orcid.org/0000-0003-2640-2879

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

La ilustración de la portada es cortesía de Sandra de la Cruz.

https://www.sandradelacruz.com/

Tesis Doctoral

Facultad de Ciencias

Departamento de Biología Molecular y Bioquímica

Modelado y explotación de redes para la caracterización y predicción en sistemas biomédicos

Realizado por

D. ANÍBAL BUENO AMORÓS

Dirigido por

DR. JUAN ANTONIO GARCÍA RANEA (Universidad de Málaga)

Departamento de Biología Molecular y Bioquímica

Programa de Doctorado: Fundamentos celulares y moleculares de los seres vivos

Málaga, 2019

Autor: Aníbal Bueno Amorós

Edita: Publicaciones y Divulgación Científica. Universidad de Málaga.

https://orcid.org/0000-0003-2640-2879

A lo largo del manuscrito se han utilizado imágenes de elaboración propia o bajo licencia Creative Commons (CC). En caso contrario se especifica la fuente de las mismas.

En los capítulos 5 y 6 las imágenes fueron extraídas de Bueno *et al.* (2016) y Bueno *et al.* (2018), respectivamente; por tratarse de capítulos centrados en dichos trabajos.

UNIVERSIDAD DE MÁLAGA

Departamento de Biología Molecular y Bioquímica

Facultad de Ciencias

El **Dr. D. Miguel Ángel Medina Torres**, Director del Departamento de Biología Molecular y Bioquímica de la Facultad de Ciencias de la Universidad de Málaga,

ACREDITA QUE la Tesis Doctoral titulada '**Modelado y explotación de redes para la caracterización y predicción en sistemas biomédicos**', que presenta **D. Aníbal Bueno Amorós** para optar al título de Doctor en Ciencias, ha sido realizada bajo la dirección del Dr. D. Juan Antonio García Ranea, Profesor Titular de Universidad del Departamento de Biología Molecular y Bioquímica de la Facultad de Ciencias de la Universidad de Málaga.

Y para que así conste a los efectos oportunos, firma el presente documento en Málaga, a 20 de Noviembre de 2018.

Fdo.: Dr. D. Miguel Ángel Medina Torres

UNIVERSIDAD DE MÁLAGA

Departamento de Biología Molecular y Bioquímica

Facultad de Ciencias

El **Dr. D. Juan Antonio García Ranea**, Profesor Titular del Departamento de Biología Molecular y Bioquímica de la Universidad de Málaga,

INFORMA QUE **D. Aníbal Bueno Amorós**, Ingeniero Superior en Informática por la Universidad de Alicante, ha realizado bajo su dirección los trabajos de investigación recogidos en el presente documento, de título '**Modelado y explotación de redes para la caracterización y predicción en sistemas biomédicos**', para optar al título de Doctor en Ciencias por la Universidad de Málaga. Así mismo, el profesor D. Juan Antonio García Ranea ha sido el tutor del doctorando durante toda su etapa predoctoral. El doctorando ha llevado a cabo la mayor parte de los experimentos aquí recogidos en las instalaciones del Departamento de Biología Molecular y Bioquímica de la Universidad de Málaga. No obstante, algunos experimentos los ha realizado en el laboratorio del Dr. Manuel Corpas (*The Genome Analysis Centre*, Norwich, Reino Unido) y en el de la Dra. Christine Orengo (*University College London*, Londres, Reino Unido).

Revisado el trabajo, estima que puede ser presentado al Tribunal que ha de evaluarlo.

Y para que así conste a los efectos oportunos, firma el presente documento en Málaga, a 20 de Noviembre de 2018.

Fdo.: Dr. D. Juan Antonio García Ranea

UNIVERSIDAD DE MÁLAGA

Departamento de Biología Molecular y Bioquímica

Facultad de Ciencias

El **Dr. D. Juan Antonio García Ranea**, Profesor Titular del Departamento de Biología Molecular y Bioquímica de la Universidad de Málaga,

AUTORIZA a **D. Aníbal Bueno Amorós** a la lectura y defensa de esta memoria de Tesis Doctoral, de título '**Modelado y explotación de redes para la caracterización y predicción en sistemas biomédicos**'.

Asimismo, el director de la Tesis informa que todas las publicaciones que avalan la presente memoria de Tesis Doctoral no han sido utilizadas en Tesis anteriores, y han sido generadas, parcial o totalmente, a partir de resultados y metodologías desarrolladas durante la realización de la Tesis Doctoral de D. Aníbal Bueno Amorós.

Y para que así conste a los efectos oportunos, firma el presente documento en Málaga, a 20 de Noviembre de 2018.

Fdo.: Dr. D. Juan Antonio García Ranea.

Índice general

Índice de figuras

Índice de tablas

Financiación

Este trabajo de Tesis Doctoral ha sido desarrollado en la Universidad de Málaga y subvencionado por la Red de Excelencia Europea *Systems Microscopy* de la Comisión Europea (EU-FP7-Systems Microscopy [258068]), el Ministerio de Economía y Competitividad del Gobierno de España con Fondos de Desarrollo Regional Europeos (SAF2012-33110) y el Gobierno de Andalucía con Fondos de Desarrollo Regional Europeos (CTS-486).

La Universidad de Málaga ha financiado la estancia predoctoral del doctorando en *University College London* (Reino Unido), incluyendo colaboraciones con *The Genome Analysis Centre*, durante el periodo comprendido entre el 1 de Abril de 2015 y el 30 de Junio de 2015, a través de una 'Ayuda para estancia en centro de investigación de calidad'.

Figura 1: **Logotipos de entidades financiadoras.**

Publicaciones y comunicaciones

Parte de los resultados y metodologías recogidas en la presente Memoria de Tesis Doctoral han dado lugar a las siguientes publicaciones y comunicaciones:

Publicaciones que avalan esta Tesis Doctoral

Bueno A, Morilla I, Diez D, Moya-Garcia AA, Lozano J, Ranea JAG. *Exploring the interactions of the RAS family in the human protein network and their potential implications in RAS-directed therapies.* Oncotarget 2016. doi:10.18632/oncotarget.12416. (Véase el capítulo 5.3.6).

Bueno A, Rodríguez-López R, Reyes-Palomares A, Rojano E, Corpas M, Nevado J, Lapunzina P, Sánchez-Jiménez F, Ranea JAG. *Phenotype-loci associations in networks of patients with rare disorders: application to assist in the diagnosis of novel clinical cases.* Eur J Hum Genet 2018; 26: 1451-1461. (Véase el capítulo 6.3.4).

Resto de publicaciones

García-Vilas JA, Morilla I, Bueno A, Martínez-Poveda B, Medina MÁ, Ranea JAG. *In silico prediction of targets for anti-angiogenesis and their in vitro evaluation confirm the involvement of SOD3 in angiogenesis.* Oncotarget 2018; 9. doi:10.18632/oncotarget.24693. (Véase el capítulo 4.4.3).

Rojano E, Seoane P, Bueno A, Perkins JR, Ranea JAG. *Revealing the Relationship Between Human Genome Regions and Pathological Phenotypes Through Network Analysis.* In: Rojas I, Ortuño F (eds) Bioinformatics and Biomedical Engineering. IWBBIO 2017. Lecture Notes in Computer Science, vol 10208. Springer, Cham. (Véase el apéndice).

Reyes-Palomares A, Bueno A, Rodríguez-López R, Medina MÁ, Sánchez-Jiménez F, Corpas M, Ranea JAG. *Systematic identification of phenotypically enriched loci using a patient network of genomic disorders.* BMC Genomics 2016; 17: 232. (Véase el apéndice).

Huerta-Cepas J, Bueno A, Dopazo J, Gabaldón T. *PhylomeDB: A database for genome-wide collections of gene phylogenies.* Nucleic Acids Res 2008; 36. (Véase el apéndice).

Comunicaciones

Póster 'Discovering the genetic signal underlying cancer cellular heterogeneity in drug repositioning strategy', en *2nd Annual Meeting on Systems Microscopy*, Lovaina. 2013. (Véase el apéndice).

Charla 'Análisis de la expansión evolutiva de la superfamilia Ras en redes de interacción de proteínas', en XV Jornadas de Biología Celular y Molecular, Universidad de Málaga. 2014.

Póster ’RAS superfamily evolutionary expansion in the human protein network interactome’, en *Advanced Lecture Course on Systems Biology (SysBio)*, Innsbruck. 2014. (Véase el apéndice).

Charla ’Búsqueda de nuevos genes implicados en la diferenciación maligna de células tumorales de la línea MCF10CA1a inducida por variación de la rigidez en la matriz extracelular’, en XVI Jornadas de Biología Celular y Molecular, Universidad de Málaga. 2015.

Póster ’Interaction network distances and connectivity measurements applied to gene prioritization in ECM stiffness regulation of breast cancer’, en *4th Annual Meeting on Systems Microscopy*, Viena. 2015. (Véase el apéndice).

Charla ’A phenotype predictor for rare diseases’, en *The BioJS conference*, Norwich. 2015.

Póster ’Using phenotype-*loci* network analysis in undiagnosed clinical cases of patients with rare genomic disorders’, en X Reunión Anual 2017 CIBERER (Centro de Investigación Biomédica en Red de Enfermedades Raras), Madrid. 2017. (Véase el apéndice).

A Darth Vader

- Alicia: ¿Podrías decirme, por favor, qué camino debo seguir para salir de aquí?

- Gato: Eso depende, en gran medida, del sitio al que quieras llegar.

- Alicia: Me da igual el sitio, sólo quiero salir de aquí.

- Gato: Si te da igual a donde ir entonces también da igual hacia donde te dirijas.

- Gato: **Siempre llegarás a alguna parte si caminas lo suficiente.**

Manuscrito desarrollado en un 87,4 % en aeropuertos, aviones y hospitales.

Agradecimientos

El *Homo sapiens* es un animal social, como el resto de primates. Y precisamente ese atributo es uno de los que le ha permitido llegar más lejos como especie. La evolución de la corteza prefrontal del cerebro en los grandes simios es la responsable de que individuos de estas especies sean capaces de comprender, analizar y trazar estrategias sobre complejos sistemas sociales. Recientes estudios afirman que el tamaño de dicha área cerebral en las diferentes especies es proporcional al tamaño de los grupos sociales en que viven.

Efectivamente, somos seres sociales y no somos sin los demás, y mucho menos en Ciencia, un área donde la cooperación y el apoyo mutuo es indispensable (como ya reconoció Isaac Newton en una carta allá por 1676: *... si he visto más lejos es porque estoy sentado sobre los hombros de gigantes ...*), formando este hecho parte de la propia idiosincrasia de esta área de la cultura humana desde tiempos inmemoriales, lo cual ha convertido a las universidades de todo el mundo en oasis y bastiones sagrados de la acumulación y transmisión del conocimiento, de manera Universal.

Y es por ello que no puedo más que comenzar con palabras de gratitud a todas las universidades que me han acogido, como estudiante, en cada una de las fases de mi formación: Universidad de Alicante, Universidad de Murcia, Universidad Internacional de Andalucía, Universidad de Málaga, Universitat de València, University College London, Universitat Autònoma de Barcelona y Universitat de Girona. Lugares en los que uno siempre se siente como en casa, arropado por el conocimiento, la evidencia, el pensamiento crítico y la razón.

Quisiera personificar el primer agradecimiento en el Dr. Daniel Ruiz Fernández, mi primer mentor científico, de la mano del cual desarrollé mi proyecto fin de carrera (en Ingeniería Informática) y comencé en un Programa de Doctorado posterior, y al cual le confesé por aquella misma época que me preocupaba que nadie me contratase por mis *piercings* y la maravillosa cresta roja que lucía como cabellera en aquel entonces. También me siento muy agradecido con el Dr. Toni Gabaldón Estevan, por haberme guiado a la publicación de mi primer artículo científico cuando apenas terminaba mi primer Máster (y él todavía estaba en Valencia).

Por supuesto mi agradecimiento al Dr. Juan Antonio García Ranea por darme la oportunidad de cumplir el sueño que he tenido casi desde niño, de trabajar en una Facultad de Ciencias en investigación; por su orientación en el campo de la Biología Molecular y Biología de Sistemas, así como por la dirección y supervisión de este trabajo durante todos estos años. Y al Dr. Manuel Corpas, por haberme acogido con la gran amabilidad que le caracteriza durante mi estancia predoctoral en *The Genome Analysis Centre*, Norwich (Reino Unido); donde me sentí valorado como profesional en un entorno cercano entre hispanos que hablaban inglés, y donde además pude sentir la pertenencia a un centro de investigación europeo y la forma de trabajar más allá del sistema universitario español, siendo una enriquecedora experiencia. En una línea muy similar extiendo este agradecimiento a la Dra. Christine Orengo por la oportunidad que me brindó de trabajar en colaboración con su excelente equipo en el *University College London*.

Al resto de mis compañeros de laboratorio, especialmente a la Dra. Beatriz Serrano Solano, la cual fue un pilar fundamental en el que apoyarme durante todos mis altibajos emocionales a lo largo del periodo en el que coincidimos en el laboratorio (además de hacerme la comida cada día -previo pago- y compartir divagaciones 'artísticas' de lo más variopintas). A la Dra. Rocío Rodríguez López (rociorp) por ser la tercera pata de este banco en el que nos hemos tenido que sentar con paciencia, en largas noches de: escritura, correcciones, desesperación y charlas políticas y sociales *online*. Al Dr. Ian Morilla, por la ayuda prestada en el área más matemática de la algoritmia aquí desarrollada, siendo quizás el gigante más alto sobre el que me subí a hombros. Agradecer también, de manera especial, al Dr. James Perkins, a David Velasco y al Dr. Pedro Seoane; por estimular la discusión

de los resultados de algunas partes de este trabajo y por su inestimable ayuda en la supervisión y traducción de los manuscritos en inglés.

También quería dar las gracias a la Dra. Francisca Sánchez Jiménez (Kika. Otra gran viajera.) y al Dr. Miguel Ángel Medina Torres (MAM) por su sabiduría, experiencia y guía espiritual (si es que eso existe en Ciencia) desde el *lado húmedo* del Departamento.

Al resto de miembros del *Dark Side* (área de bioinformática) del grupo de investigación *ProCel* (Bases Moleculares de la Proliferación Celular) con los que he coincidido y colaborado: Armando, Pedro, Almudena, Aurelio, Cristóbal, Nando y Elena.

Uno no tiene mucho tiempo para dedicar a las amistades cuando está escribiendo la Tesis Doctoral, y es por ello que conviene regarlas. Aquí mi ejercicio de goteo para aquellas que sé que de una forma u otra merecen aparecer aquí, porque realmente han contribuido a que esto salga adelante: Ángel Yorca, Andrea García, Alex Romero, Paula Olcina, Pepe Cana, Luis Miguel Cruzado y Kimberley McGrail.

A Lucía, el amor de mi vida, mi compañera, la persona que ha dado un vuelco a todo lo que conocía para demostrarme que otra forma de compartir una vida es posible y que crear un equipo unido con los mismos objetivos vitales es la clave para ser feliz y *libre*. Cariño, gracias por soportarme en la fase de escritura de este documento. Te quiero.

Y en penúltimo lugar quiero dar las gracias a lo más importante que tengo: MI FAMILIA. Esa familia que ha pasado por momentos amargos que nos han hecho estar más unidos cada vez. Mi madre, Angelines, la persona más dulce y afable que conozco, siempre preocupada por el bienestar de sus hijos, sacrificando el suyo propio hasta límites que poca gente conoce. Mamá, gracias, soy consciente de todo lo que has hecho por mí, lo que has evitado hacer, lo que has dicho o has callado, dentro de todas las duras circunstancias que hemos tenido en esta vida. Te quiero. A mi hermana Mari Ángeles, con la que tengo un vínculo muy especial, pues creo que siempre ha sabido leerme bastante bien (cosa difícil) y con la que compartí piso en la época universitaria en Alicante,

en la que aún éramos más diferentes de lo que somos ahora. Gracias hermana, y gracias también por haberme dado esas dos sobrinas, Luna e Inés, que son mi debilidad. A Elvira, la tercera en discordia, a la que quiero tanto que a veces aún nos peleamos, por aquello de seguir manteniendo las costumbres de pequeños, y con la que compartí uno de los primeros grandes viajes que hice, el que nos llevó a India y Nepal. No quiero olvidarme de mis cuñados: Pedro y Álex (al que tampoco tendría como agradecerle lo que ha hecho por nuestra familia en momentos clave). Y a mi tío Pedro y mi tía Mari Ángeles, mis segundos padres. ¡Tío aquí está mi Tesis!, deja de darme el coñazo...

Y no puedo terminar de otra forma que no sea escribiendo las últimas líneas de agradecimiento a mi padre, el cual por desgracia acertó, hace exactamente un año, cuando me dijo: *hijo, no me va a dar tiempo a ver tu Tesis terminada*. Pues no. Efectivamente. Te fuiste muy poco después de pronunciar esas palabras. Esa maldita enfermedad que a tanta gente arrastra se te llevó, pero aquí está tu legado. Quizás el trabajo aquí desarrollado aporte un granito de arena en la lucha contra esta enfermedad. Fueron tu disciplina y tu exigencia hacía mi, sin duda, las que me han llevado hasta aquí. Esto no habría sido posible sin tu influencia, de una manera u otra, en mi forma de ser, en el campo de la perseverancia y la consecución de objetivos académicos. Además, obviamente, de tu apoyo económico en mi formación. Te salió caro aquello de: *mientras viva correrán de mi cuenta todos tus estudios*; tantas matriculaciones en: carreras, másteres y doctorados que finalmente te habría salido más barato comprarme un piso... Sé que me querías, a tu manera. Y que estabas orgulloso de mí (últimamente me lo decías). Y... sinceramente, me habría gustado mucho que estuvieras hoy aquí. Aún me cuesta decirlo pero... Te quiero papá. Te echo de menos papá. Gracias papá.

Nos vemos en el camino.

Gracias a todos.

Figura 2: **Centros colaboradores en los estudios incluidos en esta Tesis Doctoral.**

El autor también ha tenido acceso a los recursos informáticos y a la asistencia de los técnicos expertos del SCBI (Centro de Supercomputación y Bioinformática) de la Universidad de Málaga.

Preámbulo

Corría el año 2005 cuando terminé de cursar Ingeniería Informática en la Universidad de Alicante. Aquel fue también el año en el que me matriculé por primera vez en un Programa de Doctorado, bajo la dirección del Dr. Daniel Ruiz Fernández, mi primer mentor científico, perteneciente al grupo de investigación *Ingeniería biomédica e informática para la salud*; sin saber que sería ni más ni menos que 14 años después cuando escribiría estas líneas con la intención de que figuren en el preámbulo de mi Tesis Doctoral. Aquel primer Programa de Doctorado tenía por nombre *Tecnologías para la Sociedad de la Información* y formaba parte del Departamento de Tecnología Informática y Computación de la mencionada universidad.

Tras una formación inicial exclusivamente enfocada en el ámbito de la Ingeniería Informática, y después de presentar como Trabajo Fin de Carrera un estudio relacionado con la biometría, tomé la decisión de virar, poco a poco, mi carrera profesional a ámbitos más relacionados con la Biología, la que era mi verdadera vocación. Por ello, tras matricularme del citado curso de Doctorado en la Universidad de Alicante, unos años después (2007) y de manera paralela me matriculé en el Máster en Bioinformática de la Universidad Internacional de Andalucía, con la intención de adentrarme en los nexos de unión entre ambas disciplinas. Otros 2 años después, en 2009, defendí mi Trabajo Fin de Máster de título *PhylomeDB*, el cual desarrollé en colaboración con el Centro de Investigación Príncipe Felipe (Valencia) y dio fruto a mi primera publicación científica: *Huerta-Cepas J, Bueno A, Dopazo J, Gabaldón T. PhylomeDB: A database for genome-wide collections of gene phylogenies. Nucleic Acids Res 2008; 36*. (Véase el apéndice).

No detuve mi formación en Biología ni en Informática. Durante los siguientes años me matriculé en el Grado en Biología de la Universidad de Murcia, mientras desarrollaba paralelamente mi propio proyecto empresarial en el área de la Informática y la Bioinformática a través de la empresa de servicios informáticos *Vértice Digital S.A.*, de la cual fui socio fundador y gerente. Fue en 2011 cuando, tras optar a una oferta de trabajo como investigador en el Departamento de Biología Molecular y Bioquímica de la Universidad de Málaga, recibí la llamada del Dr. Juan Antonio García Ranea para ofrecerme la incorporación al grupo de investigación de Bases Moleculares de la Proliferación Celular.

Durante el periodo comprendido entre el año 2011 y el 2015 desarrollé mi labor profesional como investigador contratado en el mencionado grupo, adscrito a la Universidad de Málaga. En ese tiempo elaboré los trabajos que posteriormente fueron publicados y hoy dotan de contenido y razón de ser a esta Tesis Doctoral:

Bueno A, Morilla I, Diez D, Moya-Garcia AA, Lozano J, Ranea JAG. *Exploring the interactions of the RAS family in the human protein network and their potential implications in RAS-directed therapies.* Oncotarget 2016. doi:10.18632/oncotarget.12416. (Véase el capítulo 5.3.6).

García-Vilas JA, Morilla I, Bueno A, Martínez-Poveda B, Medina MÁ, Ranea JAG. *In silico prediction of targets for anti-angiogenesis and their in vitro evaluation confirm the involvement of SOD3 in angiogenesis.* Oncotarget 2018; 9. doi:10.18632/oncotarget.24693. (Véase el capítulo 4.4.3).

Bueno A, Rodríguez-López R, Reyes-Palomares A, Rojano E, Corpas M, Nevado J, Lapunzina P, Sánchez-Jiménez F, Ranea JAG. *Phenotype-loci associations in networks of patients with rare disorders: application to assist in the diagnosis of novel clinical cases.* Eur J Hum Genet 2018; 26: 1451-1461. (Véase el capítulo 6.3.4).

Reyes-Palomares A, Bueno A, Rodríguez-López R, Medina MÁ, Sánchez-Jiménez F, Corpas M, Ranea JAG. *Systematic identification of phenotypically enriched loci using a patient network of genomic disorders.* BMC Genomics 2016; 17: 232. (Véase el apéndice).

Rojano E, Seoane P, Bueno A, Perkins JR, Ranea JAG. *Revealing the Relationship Between Human Genome Regions and Pathological Phenotypes Through Network Analysis.* In: Rojas I, Ortuño F (eds) Bioinformatics and Biomedical Engineering. IWBBIO 2017. Lecture Notes in Computer Science, vol 10208. Springer, Cham. (Véase el apéndice).

Tampoco mi actividad laboral en la universidad frenó mi formación. En 2012 obtuve mi certificado de suficiencia investigadora (DEA) con la tesina de título: *Algoritmos de análisis bioinformático de redes de proteínas 'Ras'*, Universidad de Alicante. En el año 2013 cursé el Máster en Biología Celular y Molecular, dentro del *Programa de Doctorado de Fundamentos Celulares y Moleculares de los Seres Vivos*, de la Facultad de Ciencias (Universidad de Málaga), defendiendo ese mismo año el Trabajo de Fin de Máster de título: *Expansión evolutiva de 'RAS' en el interactoma en red de proteínas*. Fue en 2014 cuando estudié el Máster en Ciencias Cognitivas de la Facultad de Informática de la Universidad de Málaga. En ese mismo año también, el Curso de Genética Médica impartido por la Universidad de Valencia. Y en 2015 el curso de Genética Humana por la Sociedad Española de Genética (Hospital de la Santa Creu i Sant Pau. Barcelona).

Durante mi periodo predoctoral colaboré con los siguientes centros: la Universidad de Kyoto, la Universidad de Alicante, Karolinska Institutet (Estocolmo), INGEMM (Instituto de Genética Médica y Molecular - Hospital La Paz), University College London y TGAC (The Genome Analysis Centre). Entre estos dos últimos centros, ubicados en Londres y Norwich (Reino Unido), realicé una estancia de investigación de 3 meses.

Fue ya en el año 2016 cuando abandoné físicamente la Universidad de Málaga, tras la finalización del proyecto europeo de la Red de Excelencia *Systems Microscopy* bajo el cual estaba contratado, para mudarme a la Universidad Autónoma de Barcelona con la finalidad de dar un giro de 180 grados a mi carrera profesional, estudiando el Máster en Periodismo de Viajes y fundando

las agencias de viajes alternativos *Camino sin Fin S.L.* y *Last Places S.L.*, lo cual me llevó de nuevo al emprendimiento, esta vez, en un sector completamente nuevo para mí. Tras convertirme en director y guía de viajes por África, comencé también a impartir clases en dicha Universidad asociadas a los viajes y la documentación de viajes, así como a realizar trabajos puntuales de Periodismo de Viajes y Fotografía, publicando reportajes y guías bajo la marca *Alt Experience*. Decidí que esto no detuviese mi carrera científica, la cual seguí desarrollando mediante diversas charlas divulgativas y artículos y que culmina con la publicación de la presente Tesis Doctoral.

Abreviaturas y acrónimos

aCGH: Siglas en inglés de 'Comparative Genomic Hybridization' (CGH), hibridación genómica comparativa. Un método citogenético para el análisis de las variaciones en el número de copias (CNVs -véase la definición en este mismo capítulo-) analizando el nivel de ploidía del ADN de una muestra en comparación con una referencia. La técnica implica el aislamiento y marcado de ADN de las dos fuentes a comparar (mediante hibridación competitiva fluorescente *in situ*), la desnaturalización e hibridación para una propagación normal de cromosomas en metafase y la comparación mediante un microscopio de fluorescencia de las intensidades, lo que se asocia a ganancias o pérdidas de material. La 'a' que se antepone al acrónimo corresponde a la fusión de la técnica CGH con el uso de *microarrays* de ADN, ganando en especificidad y permitiendo medir la variación del número de copias *locus* por *locus* con una resolución mayor (de 100 kilobases).

AUC: Siglas en inglés de 'Area Under the Curve', el área bajo la curva. En la Teoría de detección de señales, el AUC es una medida numérica de la fiabilidad de una predicción, que se basa en el cálculo del área delimitada bajo una curva ROC (véase la definición más adelante). Sus valores van desde 1 (mayor calidad de una predicción) hasta 0,5 (menor calidad). Véase el capítulo 3.3.2.

BLOSUM: Abreviatura en inglés de '**BLO**cks of Amino Acid **SU**bstitution **M**atrix', matriz de sustitución de bloques de aminoácidos. Se trata de una matriz usada para puntuar alineamientos entre secuencias de proteínas. Se estableció en 1992 a partir de un análisis de regiones conservadas

de familias de proteínas y frecuencias de aparición de los aminoácidos y las probabilidades de sustitución entre ellos.

bps: Abreviatura usada en genética para hacer referencia a pares de bases (del inglés 'base pairs'). Cada par consta de una unidad formada por 2 nucleobases unidas entre sí por enlaces de hidrógeno, dando estructura a los bloques del material genético.

chr: Abreviatura de cromosoma, del inglés '**chr**omosome'.

CNV: Siglas en inglés de 'Copy Number Variation', variación en el número de copias. Se define como un segmento de ADN cuyo número de copias es variable si se compara con el genoma de referencia.

CT: Abreviatura del algoritmo de tipo *kernel* de nombre 'Commute Time Diffusion Kernel'. Uno de los algoritmos de cálculo de distancias en red basado en probabilidades y detallado en el capítulo 3.2.1. El resultado obtenido mediante este algoritmo debe ser normalizado y es simétrico.

DAVID: Acrónimo en inglés de '**D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery'. Se trata de un recurso bioinformático disponible en Internet, que mantiene el *Laboratory of Immunopathogenesis and Bioinformatics* (EE.UU.), y que provee de interpretación funcional a listados de genes. Sus datos provienen de estudios genómicos. Disponible en: http://david.niaid.nih.gov/.

DECIPHER: Abreviatura en inglés de '**D**atabas**E** of genomi**C** var**I**ation and **P**henotype in **H**umans using **E**nsembl **R**esources'. Se trata de una base de datos interactiva disponible en Internet que incorpora herramientas para asistir en la interpretación de variantes genómicas en enfermedades raras, facilitando los diagnósticos clínicos en estos casos. Contiene datos genómicos y fenotípicos de miles de pacientes de todo el mundo con este tipo de patologías. Se encuentra accesible en: https://decipher.sanger.ac.uk/.

DIRP: Acrónimo en inglés de '**D**ivergent but **I**nteracting **R**as **P**airs', que se refiere a pares de proteínas Ras divergentes filogenéticamente pero con una cercana distancia en las redes de interacción de proteínas.

DK: Abreviatura del algoritmo de tipo *kernel* de nombre 'Laplacian Exponential Diffusion Kernel', usado para la medida de distancias en redes con base en estudios de probabilidad. En el capítulo 3.2.1 se puede consultar su funcionamiento.

GO: Acrónimo en inglés de '**G**ene **O**ntology Project'. Se trata de una iniciativa para representar computacionalmente el conocimiento sobre cómo los genes codifican la actividad biológica a nivel: funcional, celular y de tejidos. Los datos están conformados por ontologías formales estructuradas representando más de 40.000 conceptos biológicos, a los que se anotan genes. Es accesible en: http://www.geneontology.org/.

HPO: Siglas en inglés de '**H**uman **P**henotype **O**ntology', una ontología formal de fenotipos humanos formada por más de 13.000 términos que describen anomalías fenotípicas clínicas. Desarrollada por varios organismos internacionales con datos del 'Online Mendelian Inheritance in Man' (véase OMIM en este mismo capítulo) y de bibliografía médica. La ontología se encuentra accesible en: http://human-phenotype-ontology.github.io/.

HyI: Abreviatura de '**Hy**pergeometric **I**ndex' (índice hipergeométrico). Algoritmo probabilístico de medida del grado de asociación entre nodos de una misma capa en una red heterogénea *bipartita*, en función de su perfil de conexiones con una segunda capa (también aplicable a redes *tripartitas*, proyectando las 2 capas externas en la capa intermedia). El HyI calcula la probabilidad de tener un nivel de asociación igual o mayor entre 2 nodos dados de la red que el esperado por azar. En esta Tesis Doctoral es aplicado a una red *tripartita* de mutaciones-pacientes-fenotipos, para obtener conexiones significativas entre la primera y la última capa. Su fórmula se muestra en la figura 6.2.

iRef: Portal con información integrada de 10 bases de datos públicas de interacciones de proteínas: BIND, BioGRID, CORUM, DIP, HPRD, IntAct, MINT, MIPS/MPact y OPHID. Accesible en: http://irefindex.org/.

KEGG: Acrónimo en inglés de '**K**yoto **E**ncyclopedia of **G**enes and **G**enomes', Enciclopedia de Genes y Genomas de Kioto. Se trata de un conjunto de bases de datos de genomas disponible *online*, el cual anota genes y proteínas a rutas metabólicas y reacciones biológicas. Accesible en: https://www.genome.jp/kegg/.

LOO: Acrónimo en inglés de '**L**eave **O**ne **O**ut'. Se trata de un tipo de validación cruzada, técnica que permite evaluar los resultados de un método predictor estimando su precisión; en este caso haciendo uso de un conjunto de datos de los cuales ya conocemos su resultado y evaluando uno a uno el comportamiento del predictor para ellos. Véase el capítulo 3.3.1.

MSA: Siglas en inglés de '**M**ultiple **S**equence **A**lignment', alineamiento de secuencias múltiple. Se trata de un alineamiento entre 3 o más secuencias biológicas (aminoácidos o ácidos nucleicos) de una longitud similar. Del resultado y su análisis se pueden inferir datos de homología y relaciones evolutivas entre las entidades (proteínas o material genético).

NCBI: Siglas de '**N**ational **C**enter for **B**iotechnology **I**nformation', Centro Nacional para la Información Biotecnológica. Se trata de una parte de la Biblioteca Nacional de Medicina de Estados Unidos, dependiente de los Institutos Nacionales de Salud. Se encarga, entre otras cosas, de almacenar y actualizar la información referente a secuencias genómicas.

NGS: Siglas en inglés de '**N**ext **G**eneration **S**equencing', secuenciación de próxima generación (o también conocida como de alto rendimiento). Consiste en un conjunto de técnicas recientes que son capaces de paralelizar muchas instancias de secuenciación de ADN, produciendo millones a la vez, y por ello reduciendo los costos y tiempos al realizar el mismo proceso a gran escala.

OMIM: Acrónimo en inglés de '**O**nline **M**endelian **I**nheritance in **M**an', herencia mendeliana en el Hombre (*online*). Se trata de un proyecto activo desde 1966 y ahora accesible a través de In-

ternet en forma de una base de datos que cataloga todas las enfermedades humanas que se conocen y que tienen componente genético (con base en la evidencia de artículos científicos), así como su asociación a los genes identificados en cada caso. Accesible en: https://www.omim.org/.

PDB: Siglas en inglés de '**P**rotein **D**ata **B**ank', banco de datos de proteínas. Hace referencia a una base de datos sobre estructuras tridimensionales de proteínas. Los datos son generalmente obtenidos mediante cristalografía de rayos X o resonancia magnética nuclear. La base de datos se nutre del envío de esta información por parte de la comunidad científica de todo el mundo. Accesible en: https://www.rcsb.org/.

PINA: Acrónimo en inglés de '**P**rotein **I**nteraction **N**etwork **A**nalysis', análisis de red de interacción de proteínas. Se trata de una plataforma integrada para la construcción de redes de interacción entre proteínas con capacidad de: filtrado, análisis, visualización y gestión. Incorpora 6 bases de datos públicas de interacciones proteína-proteína: IntAct, MINT, BioGRID, DIP, HPRD y MIPS/MPact. Accesible en: http://cbg.garvan.unsw.edu.au/pina/.

PPI: Siglas en inglés de '**P**rotein **P**rotein **I**nteraction', interacciones proteína-proteína.

RBD: Siglas en inglés de '**R**as **B**inding **D**omain', dominios de unión a proteínas Ras.

RMN: Siglas de '**R**esonancia **M**agnética **N**uclear', una de las técnicas utilizadas para revelar estructuras moleculares.

r.m.s.d.: Siglas en inglés de '**r**oot **m**ean **s**quare **d**eviation', desviación de la media cuadrática. Medida estadística de la magnitud de una cantidad variable. Utilizada en este trabajo para medir la similitud entre las estructuras 3D de las proteínas.

RNA-Seq: *RNA sequencing*. Técnicas de *next-generation sequencing* (vease NGS en este mismo capítulo) utilizadas para revelar la presencia y cantidad de RNA en una muestra biológica en un momento dado. De esta manera se puede analizar el transcriptoma celular en el momento deseado.

ROC: Acrónimo en inglés de '**R**eceiver **O**perating **C**haracteristic'. En la Teoría de detección de señales, una curva ROC es una representación gráfica que muestra la tasa de verdaderos positivos frente a falsos positivos, en un sistema predictor. Permite evaluar la fiabilidad de las predicciones en comparación con lo esperado por azar. Véase el capítulo 3.3.2.

RWR: Siglas en inglés de '**R**andom **W**alk With **R**estart'. Algoritmo basado en probabilidades para el estudio de distancias en red. No tiene propiedades de *kernel* (no es simétrico), aunque se puede convertir en simétrico calculando la media de las distancias bidireccionales. Detallado en el capítulo 3.2.1. También existe una variante, sin función de reinicio, abreviada como **RW**.

siRNA: Siglas en inglés de '**s**mall **i**nterfering **RNA**'. Se trata de moléculas de ARN de doble hebra (complementarias), de unos 20 nucleótidos de longitud. Es un tipo de ARN interferente altamente específico para la secuencia de nucleótidos de su ARN mensajero diana, lo que le permite interferir eficazmente en la expresión del gen asociado.

SNP: Acrónimo en inglés de '**S**ingle **N**ucleotide **P**olymorphism', polimorfismo de nucleótido simple. Hace referencia a una variación en la secuencia de ADN que afecta únicamente a un nucleótido.

SOR: Acrónimo en inglés de '**S**mall **O**verlapping **R**egion', pequeña región solapante. Se trata de un área del genoma en la que coinciden 2 o más mutaciones de un conjunto de pacientes.

STRING: Acrónimo en inglés de '**S**earch **T**ool for the **R**etrieval of **In**teracting **G**enes/Proteins', herramienta de búsqueda para la recuperación de genes/proteínas interactuantes. Se trata de un recurso *web* gratuito y una base de datos de interacciones conocidas entre proteínas. Contiene información de diferentes fuentes que se actualizan con frecuencia. Accesible en: https://string-db.org/.

TCGA: Siglas en inglés de '**T**he **C**ancer **G**enome **A**tlas', el atlas del genoma del cáncer. Proyecto dedicado a catalogar los cambios biológicos a nivel molecular que son responsables de la

aparición de cáncer. Fue iniciado en Estados Unidos en 2005 y es supervisado por el *National Cancer Institute* y el *National Human Genome Research Institute*.

UniProt: Abreviatura de **Uni**versal **prot**ein resource. Repositorio central de datos sobre proteínas creado por: Swiss-Prot, TrEMBL y PIRt. Accesible en: https://www.uniprot.org/.

Estructura del documento

Esta Tesis Doctoral presenta la siguiente estructura:

- Resumen en inglés de la Tesis Doctoral.
- Introducción general con conceptos básicos transversales y el estado del arte de los temas fundamentales que componen la Tesis Doctoral.
- Hipótesis y objetivos generales y particulares de cada uno de los estudios que componen la Tesis Doctoral.
- Materiales y métodos generales relativos a los estudios que componen la Tesis Doctoral; así como otras técnicas relacionadas.
- Estudio 1: *Implementación de métodos de predicción funcional basados en redes de interacción proteína-proteína: aplicación a sistemas de diferenciación maligna de células tumorales y a la angiogénesis*. Incluye apartados propios de: Introducción, Material y métodos y Resultados y discusión.

  A partir de un estudio con células tumorales (línea MCF10CA1a, en cáncer de mama) llevado a cabo por el Instituto *Karolinska* (Estocolmo) en el cual se detectó una diferenciación maligna asociada a cambios en la rigidez de la matriz extracelular; y en base a un listado de proteínas candidatas a estar implicadas en dicho proceso tras estudios de expresión diferencial en experimentos de transcriptómica y proteómica para ambos estados (maligno y

tumoral en reposo), se detalla la metodología desarrollada para la predicción de estos nuevos genes/proteínas implicados en dicha transformación, de entre los candidatos. Finalmente dicho trabajo no fue publicado y la metodología presentada en este capítulo fue utilizada en el marco de otra investigación enfocada al descubrimiento de nuevas proteínas implicadas en angiogénesis (dentro del Departamento de Biología Molecular y Bioquímica de la Universidad de Málaga), la cual sí concluyó con éxito en una publicación, y se adjunta en este capítulo de la Tesis Doctoral.

- Estudio 2: *Exploración de la red de interacciones de la familia de proteínas quinasas RAS en humanos y análisis de su evolución funcional orientado a potenciales implicaciones terapéuticas en cáncer*. Incluye apartados propios de: Introducción, Material y métodos y Resultados y discusión.

  Análisis realizado a la familia de proteínas parálogas RAS en *Homo sapiens* con la finalidad de discernir la relación entre su evolución funcional y molecular, y el cual incluye: estudio de la red de interacciones, de las distancias filogenéticas y de los detalles estructurales de dichas proteínas; con la finalidad de abordar posibles nuevas estrategias terapéuticas para enfermedades asociadas (tales como cáncer) teniendo en cuenta los datos de conservación molecular que pueden estar influyendo en las interacciones de estas proteínas; y fruto del cual se publicó uno de los artículos que se adjunta como aval de esta Tesis Doctoral.

- Estudio 3: *Asociaciones fenotipo-loci en redes de pacientes con trastornos genéticos raros: aplicación en la asistencia al diagnóstico de nuevos casos clínicos*. Incluye apartados propios de: Introducción, Material y métodos y Resultados y discusión.

  Estudio realizado haciendo uso de la información contenida en la base de datos de enfermedades raras DECIPHER, analizando las CNVs y fenotipos asociados a los pacientes mediante técnicas matemáticas aplicadas a redes *tripartitas* (CNVs-pacientes-fenotipos) con la finalidad de obtener relaciones fenotipo-genotipo estadísticamente significativas que puedan ayudar en el proceso de diagnóstico clínico y la caracterización de estas patologías; fruto del cual se publicó otro de los artículos que se adjuntan como aval de esta Tesis Doctoral.

- Discusión general de los resultados de la Tesis Doctoral.
- Conclusiones de la Tesis Doctoral en español.
- Conclusiones de la Tesis Doctoral en inglés.
- Apéndice con publicaciones y comunicaciones científicas adicionales.

  Conference paper: Revealing the relationship between human genome regions and pathological phenotypes through network analysis. Bioinformatics and Biomedical Engineering. Lecture Notes in Computer Science. IWBBIO 2017.

  Artículo que forma parte de un trabajo de colaboración publicado como resultado de diferentes análisis llevados a cabo con los datos de DECIPHER: *Systematic identification of phenotypically enriched loci using a patient network of genomic disorders. BMC Genomics 2016*.

  Artículo publicado como resultado del Trabajo Fin de Máster (TFM) del Máster en Bioinformática de la Universidad Internacional de Andalucía: *PhylomeDB: A database for genome-wide collections of gene phylogenies. Nucleic Acids Research 2008*.

  Pósteres presentados a congresos científicos como parte de la difusión de los trabajos realizados a lo largo de esta Tesis Doctoral.

No, no lo intentes. Hazlo o no lo hagas, pero no lo intentes.

Maestro Yoda

Summary

General introduction

The new technologies that have emerged in recent years, applied to biology, have made it possible to obtain the complete genomes of a multitude of animal and plant species, as well as their: transcriptomics, interactomics, metabolomics, etc. And thanks to this information and the recently developed genomic and proteomic analysis technology, great advances have been made in the characterization and treatment of various diseases. However, the gap between the generation of information and the acquisition of knowledge through it continues to increase. The understanding of how the different molecular levels are related to carry out all the vital functions remains the greatest challenge in the life sciences (it is estimated that only 10 % of all interactions between proteins in *Homo sapiens* are currently known [1]).

New scenarios require new tools and new points of view for analysis. Therefore, we are witnessing (since the beginning of the century) the birth of a new scientific speciality: Molecular Systems Biology; which approaches the biological analyses holistically instead of in a reductionist way, and this allows obtaining, from complex systems, emergent systemic conclusions. In the last decade the vision of different areas of scientific knowledge as complex systems has conceptually transformed the way of approaching their study, the methodologies and the conclusions of the experiments. This paradigm shift allows research from a non-linear point of view, avoiding the aforementioned reduc-

tionism that in some areas has led to the attainment of erroneous conclusions or to the impossibility of detecting the emergent properties of the system [2].

The present work is included within this new paradigm, being developed fundamentally in the discipline of Molecular Systems Biology (or Systems Medicine in some cases), modeling different types of complex biological systems in association networks and analyzing their properties for the prediction of new functional relationships. Mainly, the studies presented here have been focused on the analysis of:

- 1) Phylogenetic relationships.
- 2) Protein interaction networks.
- 3) Biomedical networks of associations between: mutations, patients and pathological phenotypes.

Mostly, the studies included in this Doctoral Thesis have applications in therapeutic derivatives or clinical or genetic diagnosis. The prestigious scientist in biological networks Albert-László Barabási highlights the fact that Systems Medicine is becoming an essential approach to establish molecular relationships between different pathological phenotypes, identify new genes in diseases and find the effect of mutations and their association with diseases [3]. Systems Medicine (closely related to Systems Biology) makes use of network theory and has been able to determine some general properties inherent to biological networks [4, 5].

The analyses of interaction networks between biomedical entities have been applied to different areas: from the protein interaction networks and their phylogenetic relationships, passing through the patients and the network they comprise, until reaching the area of diagnostic medicine, relating clinical phenotypes in different syndromes.

Phylogenetic relationships

Phylogenetics is the science that studies the evolutionary history of organisms, more specifically of the genes of these organisms. At the molecular level, the construction of phylogenetic trees allows the visual and numerical interpretation of gradual differences between molecules of the same organism (paralogs) or of different organisms (orthologs). In this way is possible to establish degrees of similarity in groups of molecules (either DNA or proteins). From their sequences the similarities and differences are calculated mathematically and, together with the known evolutionary data, a 'molecular evolutionary tree' can be reconstructed and information extracted from it.

Throughout the evolutionary history, in a genome, processes of *gene duplication* are produced. This event occurs when, derived from a failure of replication when copying the genome in a cell division, the same gene has been copied twice.

When this duplication occurs, 2 fragments of the genome coexist that may encode the same protein. Depending on the mutations that take place in the original gene and in the duplicated gene, the sequences will change, in a process called *divergence* (and in the same way the protein they encode), giving rise to proteins with a common origin but with a different sequence and functionality. An analysis of the sequences of both proteins (or of the sequences of both genes) provides a view of the divergence between both and allows us to quantify their degree of phylogenetic *distance*. Applying this to whole protein families (or gene families), the mentioned evolutionary history -at a molecular level- can be reconstructed between them.

Nowadays, with the new bioinformatic techniques, the amount of phylogenetic data that is generated and stored in online databases is huge and normally available to the scientific community.

In the context of the studies presented here, phylogenetic trees were used as a way to catalog and measure the evolutionary distances in the RAS protein family and their subsequent comparison with the functional relationships in the human interactome.

Protein interaction networks

Protein interaction networks try to reconstruct, from different available data sources, all the interactions that occur in living organisms between proteins: the interactome. They represent proteins as nodes in the network and its interactions as edges. The study of protein interaction networks plays an increasingly important role in the understanding of cellular mechanisms and diseases. The first analysis of interactions for the reconstruction of metabolic pathways, as we know it, was made in the 40s, but it was at the end of the 90s when computing technologies promoted exponential growth in this type of analysis. Currently, there is a significant amount of databases available with this kind of information and there are more and more techniques that are applied on a large scale to analyze the huge amount of data they contain.

One of the main challenges of Molecular Systems Biology is the reconstruction of all the interactions between proteins that take place in cells (the so-called interactome). Advances in this sense would help to understand the internal machinery of the cell, allowing the design of drugs or the characterization of the mechanisms associated with cellular diseases.

In this work, several databases were analyzed and a selection was made for their study, based on: the reliability, the origin of the information and the coverage with the studied data, in each case. The corresponding network models were generated and analysis methods were implemented for their exploitation, with the purpose of measuring *distances* between protein pairs and being able to use this information in two of the studies presented in this Doctoral Thesis: i) constructing predictors for the identification of new proteins involved in molecular processes and ii) comparing evolutionary differences with functional differences within a protein family.

Biomedical networks of associations between: mutations, patients with rare diseases and pathological phenotypes

The integration and large-scale comparison of phenotypes and genotypes of patients with genetic disorders is essential for their diagnosis. Therefore, progress in the characterization of the genetic regions and the molecular mechanisms that control phenotypic expression is crucial.

We define a rare disease as the one affecting a small proportion of the population (according to the European Union: 1 in every 2,000 people [6]), and they are many times associated with CNVs. A CNV (Copy Number Variation) is a DNA segment (at least 1kb in size) whose number of copies in the genome of an individual is variable with respect to the reference genome. It is a structural anomaly in the form of deletion, duplication or translocation, either inherited or by spontaneous occurrence (*de novo*), with potential pathological effects [7]. These pathological effects, or phenotypes, refer to the set of observable characteristics that an individual presents as a result of the interaction between his genotype and the environment [8]. Phenotyping matters [9], and has an importance that increases as we deepen our knowledge about genetic disorders, therefore the standardization of phenotypic terms, the universalization of their way of being codified and their ontological classification are of great importance for the automation of analyses in Systems Medicine [10].

The full identification of the phenotypic consequences of CNVs remains a challenge. It is necessary to take into account a large number of molecular and genetic mechanisms to determine the relationship between CNVs and the phenotypes of an individual. This is even more complicated when applied to the study of rare diseases (as has been done in this Doctoral Thesis), due to the extremely low number of patients per disease and the lack of information and characterization of these pathologies.

In this work a *tripartite* network (genotypes-patients-phenotypes) was constructed analyzing data from patients with rare diseases in order to detect genotype-phenotype relationships that could help in the clinical diagnosis and the characterization of this type of pathologies.

Hypothesis and objectives

From the foregoing, this Doctoral Thesis has the general hypothesis that the modeling of biological systems in association networks, like the ones built in this work, significantly helps the characterization and understanding of these systems in different areas: molecular (at the level of the function of proteins or genes) and phenotypic (at the level of pathologies or symptoms); also its exploitation helps to make predictions that allow optimizing laboratory experimentation for the identification of new functional proteins, as well as to assist and guide clinical diagnoses.

Dividing the main hypothesis in each of the three different research lines of which this Doctoral Thesis is composed, we have a more specific sub-hypotheses for each of the studies carried out:

1. The functional role of new proteins not yet characterized, in certain molecular systems, can be predicted through the analysis of their network relationships with proteins already known to be part of these given systems.

2. There is a relationship between the phylogenetic evolution of pairs of sequences of RAS paralogous proteins and their location in the human interactome.

3. The integration into heterogeneous association networks of genetic data (mutations) and symptoms (phenotypes) of thousands of patients with rare genomic disorders makes it possible to systematically identify new genotype-phenotype significant relationships.

In summary, this Doctoral Thesis aims to highlight the application of methodologies specific to Molecular Systems Biology and Systems Medicine in clinical and pharmacological research. The following objectives are established in order to verify the hypotheses:

The general objective is the construction of precise networks that model biological associations and the implementation of algorithms that allow their exploitation for the study of diverse systems at the molecular and phenotypic levels.

This objective is divided into the following subgoals:

1) To build a predictor, based on the mathematical analysis of protein interaction networks, that allows to statistically prioritize those proteins with more probabilities of being involved in a given molecular system, in order to characterize them functionally.

2) To use the interaction networks between proteins and the phylogenetic trees of the RAS paralogous proteins to study the relationship between their molecular and functional evolution in the human interactome.

3) To build a *tripartite* network based on the most complete information possible of patients suffering from rare genomic disorders. This network will contain the layers: CNVs (mutations), patients and phenotypes (organized hierarchically in an ontology). And mathematically analyze said network in order to obtain valid predictions about new CNV-phenotype relationships with the aim of assisting the clinical diagnosis in this kind of patients.

Studies included in this Doctoral Thesis

Three studies are part of this Doctoral Thesis, each of them in a different biological analysis area, which are presented in separate chapters with their corresponding sections.

Interaction network analysis and connectivity measurements with ROC validation applied to gene priorization in ECM stiffness regulation of breast cancer and angiogenesis

Introduction

A methodology was implemented to predict new genes/proteins involved in the biological systems responsible for the transformation to a malignant phenotype of cells MCF10CA1a (breast cancer) when there is a change in stiffness in the extracellular matrix [11–13]. This biological problem arose in the context of a collaboration within the **Systems Microscopy Network of Excellence** project. The bioinformatic analysis started from several **reference sets** (proteins known to be part of specific molecular systems associated with this process). After a systemic analysis of protein interaction networks by means of probabilistic algorithms (*kernels*), and the validation of the methodology by using the Leave One Out (LOO) cross validation technique and plotting the corresponding ROC (Receiver Operating Charasteristic) curves and their associated AUCs (Areas Under the Curve), the predictor was able to elucidate the probability that any of the candidate proteins were involved in any of the molecular systems.

This biocomputational approach was also used to predict new potential targets of anti-angiogenic therapies, by attempting to uncover new proteins involved in this molecular process. Finally, some of the predictions were validated *in vitro*, *ex vivo* and *in vivo* [14].

Methods

According to the cellular process described in the introduction, the molecular systems of study were first defined and a list of proteins that were known to be part of these specific biological processes created. This set is known as **reference set** and was built based on the bibliography [13, 15, 16], manual reviews of experts and public databases: on functionality (DAVID [17, 18]), metabolic

pathways (KEGG [19]) and cellular processes (GO [20]). These cellular systems were: cell-cell adhesion, cytoskeletal regulation, focal adhesion, Hippo signaling pathway, mechanical regulation of the nucleus and mechanotransduction of oncogenic signaling.

After that, 5 different **models of the human interactome**, in the form of protein interaction networks, were analyzed: iRef [21] (which provides an index of interactions between proteins from the query to several databases), PINA [22] (which integrates information from 6 different databases, creating a complete and non-redundant index), STRING Experimental (which exclusively includes experimentally validated interactions) [23], STRING Textmining [23] (with statistically relevant co-occurrences of gene/protein names in scientific texts) and Reactome Pathways [24] (which includes information obtained from metabolic pathways and reactions).

The analyses were carried out applying, to the matrices that represent the networks, **2 different *kernel* algorithms: Commute Time Kernel** (CT) [$K = L^{+}$] and **Exponential Laplacian Diffusion Kernel** (DK) [K = exp(-βL)]; in order to calculate the distances between the proteins in the interaction networks. After this, for each protein of the interactome, the arithmetic mean between it and the set of all the proteins that were part of the reference set was calculated, obtaining an overall score of its functional proximity to the studied system.

Through these different models of protein-protein interaction networks and the association metrics used, a specific set of predictors was constructed for each of the molecular systems mentioned above. That is, for each of the 6 systems, 10 predictors were generated: using 5 different interaction networks (interactome models) and 2 analysis methods, that is, 10 different combinations.

The next step was to evaluate the performance of each predictor in each system by using the **Leave One Out (LOO) cross validation technique and plotting the corresponding ROC (Receiver Operating Charasteristic) curves**. After this, making use of the AUC (Area Under the Curve) values, the combination of the protein interaction network and the algorithm that had the best performance in each system was selected.

In this way, the predictor is ready to be used, allowing us to statistically prioritize those proteins most likely to be involved in a given molecular system. These final results are already valid to guide prioritization (among the **candidate genes/proteins**) when performing future *in vitro* experimentation for confirming the involvement of a new protein/gene in a molecular process, and maybe for opening new paths in pharmacological research.

One of the properties of the methodology described here is its cross-sectional nature, since it is applicable to predict protein functional association to many other molecular systems or processes. Therefore, an extrapolation of the method was made to an investigation about angiogenesis, with the purpose of discovering new proteins involved in this process.

In this case, the **reference set** was formed by 116 proteins involved in angiogenesis, according to the expert knowledge and the available bibliography. Protein interaction networks were integrated into a single model, mainly through the fusion of the following sources: CODA (functional predictions based on data from protein domain matches) [25], GECO (functional similarity derived from gene expression data) [26] and HIPPIE (interaction homology relationships) [27]. The network analysis algorithm used was the RWR (Random Walk with Restart).

After the construction of the predictor and its validation by LOO and ROC, an ordered list was obtained by statistical significance, formed by 19,618 candidate proteins. The first 300 were analyzed -by experts- and 7 of them selected to carry out *in vitro* experiments. This decision was made based on the feasibility of the experimental design and the costs associated with it.

Finally, *in vitro* and *ex vivo* assays were carried out using gene silencing and blocking through specific antibodies. Additional *in vivo* experiments were also performed.

Results

It was observed that there were **2 combinations of interaction network and algorithm that offered the best performances** as predictors of the different molecular systems involved in the transformation to a malignant phenotype of cells MCF10CA1: the combination of CT algorithm with

iRef interaction database [21] and DK algorithm together with STRING Experimental [23]. Three of the systems had better performance with the first combination and another 3 with the second one. After applying the selected predictors to the different molecular systems, **the ROC curves showed a very good performance** in all of them. Therefore, reliable results can be expected.

When applying the described methodology to the molecular process of angiogenesis, it was found that from 7 predicted candidates, *in vitro* experimental tests clearly showed that *superoxide dismutase 3* silencing or blocking with specific antibodies inhibits both key steps of angiogenesis: endothelial cell migration and differentiation to tubular-like structures. Furthermore, angiogenesis was also inhibited in *ex vivo* and *in vivo* assays when blocking SOD3. Therefore, ***superoxide dismutase 3*** **is confirmed as a promising target for anti-angiogenic therapy, demonstrating experimentally the convenience of using the implemented *in silico* predictor**, developed in this work, for selecting the potential targets for the experiments, thus saving time and costs.

Exploring the interactions of the RAS family in the human protein network and their potential implications in RAS-directed therapies

Introduction

The RAS protein family is a set of small GTPases that behave like binary switches by alternating their activation state from GTP-bound (active) to GDP-bound (inactive). In higher eukaryotes these proteins are involved in signal transduction pathways controlling a diverse array of essential cellular functions, such as growth, differentiation and survival [28]. With the exception of a few well-studied protein models, the precise functions of the thirty-five human *RAS* paralogs and their relation in terms of sequence conservation, gene expression and protein-protein interactions remains poorly understood [29].

Of clinical relevance, up to 30 % of all human tumors present oncogenic mutations in members of the RAS family which often contribute to tumor pathogenesis by overactivating the Raf/MEK/ERK pathway [30, 31]. *KRAS* is the most frequently mutated *RAS* gene, accounting for up to 20 % of all tumors. Oncogenic RAS mutations are predominantly found in residues G12, G13 and Q61, impairing the intrinsic GTP hydrolysis and therefore, rendering RAS proteins in a permanent GTP-bound, active state.

While the 3 prototypical RAS proteins (*KRAS*, *HRAS* and *NRAS*) had been extensively characterized, much less is known about the remaining RAS paralogs in either healthy or tumor tissues. In this work, a study of the relationship between phylogenetic distances of RAS paralogs and their associations in the human protein interaction network has been carried out. To this end, it has been implemented a comparative sequence analysis to find conserved amino acid positions between divergent RAS-protein pairs that preserve protein interaction network proximities in the human interactome (Divergent but Interacting Ras Pairs -DIRP-). The hypothesis is that these positions may help maintain functionally important protein interactions common to both paralogs in DIRP pairs

resulting in close network proximity. These positions are then mapped onto different RAS complexes using their 3D structural information in order to determine their connection to RAS protein binding sites.

The results obtained add a new perspective to the generally accepted idea that the interactions between paralogous proteins diverge with their sequence [32–34] and shed some light on the largely unknown role of the human RAS interaction network. Furthermore, the findings broaden the current perspective on the putative role of paralogous genes in the development and adaptation of functional and pathological RAS signaling networks. In addition, important conclusions can be drawn out of the conserved positions in the DIRPs regarding their potential functional relevance for the design and development of new Ras inhibitors.

Methods

To answer the questions regarding the relationship between the protein interaction functionality and form of evolution of the RAS protein family, the data on phylogenetic distances were extracted and compared with the Ras protein-protein distances in the human interactome, using different analysis methods, for studying correlations.

Phylogenetic trees of the Ras family. The phylogenetic trees for the 35 human Ras paralogous proteins used in this work were part of the dataset that was obtained in Diez *et al.* [29]. These original trees were the product of an exhaustive and accurate search for all the encoding genes in the Ras protein families across 24 eukaryotic species. Ras human sequences were obtained from Uniprot and were aligned with their orthologs using ClustalW [35]. Finally, phylogenetic trees were constructed by Neighbor-Joining method implemented using the software Quicktree [36]. Tree topology reliability was assessed with the bootstrap method using 1000 replications.

Protein-protein interaction networks data. The two protein-protein interaction networks used in this work were constructed using the following human datasets: PINA and STRING [22, 23]. STRING describes 263,666 interactions between 14,732 proteins from the integration of: BIND, DIP, GRID, HPRD, IntAct, MINT and PID databases [37–42]. PINA includes 108,477 unique

interactions between 15,450 different proteins collected from 6 publicly available and manually curated databases: IntAct, MINT, BioGRID, DIP, HPRD and MIPS/MPact [43, 44]. Only direct physical interactions were used in this study, avoiding both data derived from phylogenetic studies (preventing tautologies in the results when comparing with tree distances) and interactions obtained by textmining processes.

Pairwise distances in the phylogenetic trees. Phylogenetic pairwise distances were calculated using the algorithm described by Pazos *et al.* [45], which uses protein tree files in the Newick Standard format as input and returns the numeric distance value for each pair.

Pairwise distances in PPI networks. RAS proteins were mapped onto the PPI networks and highly connected nodes (those with 300 or more connections) were removed, since these promiscuous hubs introduce noise in distance calculations, as shown by Hériché *et al.* [46]. Out of the various algorithms tried, the **Exponential Laplacian Diffusion Kernel** (DK) [K = exp(-βL)] and the **Commute Time Kernel** (CT) [$K = L^{+}$], were the ones that best fitted our purposes. Thus the pairwise protein distances within the networks were calculated using these methods. These are based on a calculation of the probability of association of node pairs in the network using different statistical approaches for mathematically representing the network flow.

Statistical comparison between phylogenetic distances and PPI network matrices and their plot representations were performed using the computational software R [47].

Selection of the Divergent but Interacting RAS Pairs. To select divergent sequence pairs a maximum identity threshold of 45 % was defined. This value was based on the BLOSUM 45 matrix [48], which was designed to weight amino acid substitutions between highly divergent sequences. And to establish significant closeness between proteins in the interaction networks, a second threshold was set based on random distributions of the DK and CT distance values. For each dataset and algorithm used, this threshold was estimated accordingly to a statistical p-value = 0.05. Finally, those pairs with sequence identity $\leq$ 45 % and DK and CT values $\geq$ [threshold p-value 0.05] were used to select the final set of DIRP (Distant but Interacting Ras Pairs).

Multiple sequence alignment and measurement of amino acid conservation. A Multiple Sequence Alignment (MSA) of all Ras sequences was employed to assess amino acid conservation between protein pairs, using the BLOSUM 45 amino acid substitution matrix, based on the fact that this matrix was originally designed to compare highly divergent sequences with up to 45 % identity, a condition that the dataset mostly fulfilled. Only those amino acids that aligned with the HRas sequence were used for the analysis of conservation. HRas was selected as a template for being the most studied protein in the family and one of the main pharmacological targets. For each amino acid position in the MSA, a p-value was calculated and used as a threshold to select the significantly conserved amino acids in the DIRP (those with p-value ≤ 0.01), based on random models.

Acquisition and processing of Ras complexes structural data. All known interaction complexes of human Ras proteins were downloaded from the Protein Data Bank (PDB) [49]. For each functional group, the Ras interaction surface was determined by computing the difference in the solvent accessible surface area of Ras amino acids between the complex and unbounded states, using the DSSP software [50]. And a final study was carried out to determine which of the significantly conserved amino acids in the DIRP were part of the interaction surfaces of Ras having special importance in the maintenance of their functional context (despite their phylogenetic distance). This can be useful for the search of therapeutic targets.

Results

Phylogenetic and network distance relationships of human RAS paralogs. RAS paralogs tended to be closely associated in the interactome when they were phylogenetically close and to increase their distance as they diverged. The inverse correlation between sequence similarity and the phylogenetic distance of Ras protein pairs is consistent with an evolutionary model by which recently duplicated genes share the same context of interactions. Thus, as sequences diverge by accumulation of mutations, they move away from each other in the interactome. However, the results show that some of the distant duplicated genes keep a similar protein-protein interaction context, suggesting that there is more to this model.

Identification of divergent Ras paralog pairs located close in the PPI network. The set of Distant but Interacting Ras Pairs (DIRPs) was selected out of all RAS pairs based on two statistical thresholds of significance: i) significant sequence divergence between proteins in the pair and ii) significant closeness in the protein interactome. Being -the DIRPs- between 20 % and 30 % of the initial pairs in the different networks.

Searching for conserved positions in Divergent but Interacting RAS Pairs (DIRPs). The described methodology allowed to identify positions significantly and specifically conserved in the DIRP dataset compared against both, the random model and the whole background MSA. With this approach a total of 22 positions were selected (p-value ≤ 0.01). The absence of Ras protein pairs that are similar in sequence but separated from each other in the interactome contrasts with the abundance of highly divergent Ras pairs close in the network. This suggests that a protein needs to accumulate many neutral and adaptive point mutations in order to get new interacting partners, whilst it can maintain its interaction context through a few key conserved positions.

Relationship between the DIRP conserved positions and the Ras protein binding regions. Out of the 22 DIRP specific positions identified, 15 (68 %) are directly involved in one or more binding regions and are located in some of the functional regions identified in Ras proteins. Another 4 are surrounded by 2 consecutive interacting positions in the amino acid sequence. Considering that these last positions may also be involved in Ras protein-protein interactions, we can conclude that 86 % of the DIRP specific positions participate in the interactions of Ras with other proteins. The remaining 3 were not related to any known interaction site in this analysis. These results indicate that DIRP specific positions are important to establish interactions between Ras and its partners and therefore their conservation can be an important factor in maintaining these phylogenetically distant Ras paralogs close in the interactome.

This study broadens our understanding of the human RAS signaling network and stresses the potential relevance of cross-talking, which should be taken into account when considering the inhibitory activity of drugs targeting specific Ras oncoproteins.

Phenotype-*loci* associations in networks of patients with rare disorders: application to assist in the diagnosis of novel clinical cases

Introduction

It is now widely accepted that deep phenotyping [51] and genotypic characterization of patients accelerates the identification of new genetic diseases as well as improves our understanding of the molecular basis of human genetic pathologies [10, 52, 53]. However, the accurate diagnosis of many genetic disorders becomes more complicated when patients show complex phenotypic profiles [54], when several genomic syndromes share clinical features among them, or when rare genetic aberrations affect an extremely low number of patients, as in rare diseases. Hence, key challenges for clinicians include the interpretation or classification of novel/extremely rare variants and the understanding of the phenotypic consequences of these genetic variations. A 'genotype first' approach, in which patients are classified by a similar genomic rearrangement before a common clinical presentation is observed, has proven to be successful in characterizing the growing list of microdeletion/microduplication syndromes in the last times [55, 56].

CNVs may be the cause of many disorders (such as schizophrenia, Crohn's disease or autism) and their identification and analysis are used for the diagnosis and characterization of many chromosomal syndromes [57–59]. Nowadays, the complete identification of the phenotypic consequences of a given CNV remains challenging. Thus, it is imperative that new significant advances are achieved in the characterization of the genetic regions and molecular mechanisms controlling phenotypic expression.

To help with the characterization of molecular relationships between different phenotypes and microvariants, principles of Network Medicine [3–5, 60, 61] were applied in this work to find the phenotypic consequences of variants and their association with diseases. To this end, a computational approach was developed via *tripartite* networks made of three types of nodes: variants (CNVs),

patients and phenotypes. The DECIPHER database, a global repository of clinical patient data, was used as a data source for a systematic analysis and characterization of CNVs that are likely to affect function [62]. DECIPHER is a valuable resource that offers the phenotype and genotype records of a sizable number of patients with low prevalent genomic disorders, collected from more than two hundred institutions from around the world [7, 62, 63]. Most patients with CNVs in the DECIPHER database correspond to paediatric disorders related to developmental delay, mental retardation or congenital structural anomalies [54, 64]. Along with CNVs, DECIPHER provides the pathological phenotypic profiles of the patients. This information is stored using a normalized vocabulary of phenotypes: the Human Phenotype Ontology (HPO) [10], that facilitates the analysis and comparison between patient symptomatologies. In order to study the genotype-phenotype relationships in this dataset, the associations in the *tripartite* networks were exploited using the subset of patients presenting *de novo* CNVs in DECIPHER, identifying significant associations between mutated regions and pathological phenotypes. These phenotype-*locus* associations have been used to assess the potential of the network approach for assisting the diagnosis of novel uncharacterized rare cases in clinical practice. This approach shows the potential of integrating information from thousands of characterised cases to identify novel genotype-phenotype patterns in rare and isolated cases with very scarce information to compare with.

Methods

Source of datasets for building the networks. The annotated *de novo* CNVs garnered from DECIPHER patients with rare genomic disorders [7, 62, 63] and their available HPO terms (pathological phenotypes) [10, 52, 53] were used to build a *tripartite* network formed of three layers: genotypes (by CNVs), patients and phenotypes (by HPO terms). The DECIPHER 2014 dataset version was used because this version did not include data from the patients used here as novel clinical cases, as they were included in later versions, allowing us to test the feasibility of the network approach presented in this work. In addition, we focused only on *de novo* mutations, as they are more likely to be associated with pathological phenotypes [65]. As HPO is organized as a hie-

rarchical tree, each patient was associated, in the network model, to his/her specific HPO terms (children) and all the parental terms above them in the HPO tree.

Generating the network model. CNVs were divided into deletions and duplications, as it has been previously observed that they may have different effects when affecting the same region [55, 66]. The deletions subnetwork included 2,436 *de novo* CNVs from 2,301 patients and 1,795 HPO phenotypes. Duplications subnetwork was formed by 1,114 *de novo* CNVs from 1,013 patients, including 1,160 HPO terms. The deletions and duplications subnetworks generated 45,361 unique HPO term-patient / 30,038 *loci*-patient associations and 17,010 unique HPO term-patient / 10,888 *loci*-patient associations, respectively.

Phenotype-genotype association measure calculation. Hypergeometric Index (HyI) was used to measure the degree of association between HPO terms and *loci* through patient nodes in the *tripartite* networks. The HyI yields the minus log-transformed probability of having an equal or greater level of interaction between a given phenotype-*locus* pair than the one expected by chance [67]. The significance of the association increases with the HyI value, since the lower the probability of the observed phenotype-*locus* association to be due to random the higher the HyI score value. In order to establish a significance threshold, for this study, HyI values ≥ 2.0 were considered as significant HPO phenotype-*loci* associations (p-value ≤ 0.01). The system calculated the HyI association score for 600,234 different HPO term-*locus* pairs using the deletions subnetwork, and 175,956 using the duplications subnetwork.

Ranking putative Phenotype/CNV associations in novel uncharacterized clinical cases. Finally, the system is able to identify phenotype-genotype associations, ranked by their HyI value, for new patients that were not included in the DECIPHER dataset and in this way assist in the clinical diagnosis of patients with rare diseases.

Clinical case datasets used for testing the network approach. The methodology was tested with two different cohorts of patients from INGEMM (Institute of Medical and Molecular Genetics, Hospital La Paz, Madrid, Spain): 1) single clinical cases: The first set of cases corresponds

to a cohort of 293 patients (unpublished data) showing 519 genetic aberrations (312 deletions, 155 duplications and 52 complex rearrangements), which were identified using oligonucleotide array CGH or SNParrays. These patients were mainly referred to the clinics due to: intellectual disability, congenital malformations and autistic spectrum disorder. 2) A group of patients sharing phenotype and genotype, describing a new microdeletion/microduplication syndrome: The second group of cases used was based on a specific syndrome characterization study, carried out by Nevado *et al.* [55], including 13 unrelated patients (with a total of 15 genomic rearrangements, distributed into 13 deletions and 2 duplications, located at the 19p13.3 genomic region). 11 patients had deletions and 2 of them a single duplication. The aCGH analysis together with clinical records showed that these patients shared phenotypic and genotypic features representing a novel interstitial microdeletion/microduplication syndrome [55]. Common features consist of: abnormal head circumference (macrocephaly for the deletions and microcephaly for the duplications), intellectual disability, developmental delay, hypotonia, speech delay and some dysmorphic features.

Results

Application of the networks analyses to novel clinical cases. For 258 out of the 293 clinical cases (88 %) the system found at least an overlapping CNV with a pathological *locus*, and for each pathological *locus* a list of HPO phenotypes sorted by their HyI value was provided. The resulting ranked list for each pathological *locus* only included HPO terms with a HyI value ≥ 2.0 (p-value ≤ 0.01). A total of 381 out of the 1,489 HPO terms (26 %) diagnosed by clinicians were also identified by the system associated to a patient's CNV in the *de novo* deletions subnetwork, and 252 out of 609 (41 %) in the *de novo* duplications subnetwork. On the other hand, a total of 521 and 376 non-diagnosed HPO phenotypes, for deletions and duplications respectively, were identified by the method to be associated with disorders in the clinical cases; suggesting the need to carry out some additional clinical tests to confirm or discard these phenotypes in the patients. These results indicate that this novel approach could be extensively used for differential diagnosis of novel clinical cases in order to find those phenotypes associated with single CNVs through comparison to the entire patient information integrated in the network generated from DECIPHER.

Application of the methodology to a set of patients who share a novel non-recurrent microdeletion/microduplication syndrome. All these patients share a number of phenotypes related to a similar CNV rearrangement (deletions and duplications). The results show that the systematic approach was able to identify 128 out of the 178 diagnosed phenotype-patient associations for this syndrome (72 %). Where the 21 % (37 phenotype-patient associations) showed significant HyI values (HyI $\geq$ 2.0; p-value $\leq$ 0.01), and the remaining 51 % (91 phenotype-patient associations) lower HyI values. There is a set of phenotypes diagnosed in most patients with this syndrome that were also recurrently found by the systematic approach, with significant HyI values. The system also found phenotypes associated with these CNVs in 46 % of the patients with this syndrome that had not been reported in the patients' clinical records, such as: 'Abnormality of the kidney', 'Abnormality of the penis' and 'Abnormality of connective tissue'. In a retrospective review (carried out after the application of the method to check the predictions directly in patients) of 38 of these patients with 19p13.3 microdeletions, renal anomalies were found in 26.31 % of them, anomalies of the sexual organs in 21.05 %, however there were no known cases of abnormality of the connective tissue. These results support the potential of the system to assist clinical diagnosis.

The systematic approach implemented in this work is able to better define the relationships between phenotypes and specific *loci*, by exploiting large-scale association networks of phenotypes and genotypes in thousands of rare disease patients. The application of the described methodology facilitates the diagnosis of novel clinical cases, ranking phenotypes by *locus* specificity and reporting putative new clinical features that may suggest additional clinical follow-ups. In this work, the proof of concept developed over a set of novel clinical cases demonstrates that this network-based methodology might help improve the precision of patient clinical records and the characterization of rare syndromes.

Conclusions

1) The development of predictors based on the exploitation of protein interaction networks using *kernel* analysis algorithms, validated by LOO and ROC curves, is an effective method to prioritize proteins potentially involved in a molecular mechanism of reference, with possible therapeutic implications when pathological processes are studied.

2) The comparison of phylogenetic distances, between protein pairs of the RAS family of paralogs, and their distances in the interactome (expressed by means of protein networks) is a valid methodology to study the relationships between their molecular and functional evolution, being able to obtain, in certain cases, details of the molecular/structural changes that in turn determine differences or similarities in the context of interactions of RAS paralogs with third proteins.

3) The construction of a robust network, with 3 levels (*tripartite*), which includes: genomic mutations (CNVs), patients with rare genomic disorders and phenotypes categorized in a formal ontology (HPO); and its analysis by means of specific algorithms to study network associations is a procedure capable of identifying significant relationships between mutated regions and phenotypes, providing a useful tool for assisting the clinical diagnosis of patients with such pathologies.

The general conclusion of this Doctoral Thesis is that the modeling of biological systems in the form of association networks allows advances in the characterization of such systems, and the mathematical exploitation of these models is useful for making predictions, allowing us to optimize laboratory experimentation for the identification of new functional proteins. In the same way, analyses of biomedical association networks, based on patient data, allow the construction of tools capable of assisting and guiding clinical diagnoses.

Tesis Doctoral

Modelado y explotación de redes para la caracterización y predicción en sistemas biomédicos

ANÍBAL BUENO AMORÓS

Departamento de Biología Molecular y Bioquímica
Universidad de Málaga

Capítulo 1

Introducción general

Nada tiene sentido en Biología si no es a la luz de la evolución.

Theodosius Dobzhansky

Las nuevas tecnologías surgidas en los últimos años, aplicadas a la Biología, han permitido obtener los genomas completos de multitud de especies animales y vegetales, así como su: transcriptómica, interactómica, metabolómica, etc[1]. Y gracias a esta información y a los procesos de análisis genómicos y proteómicos en torno a ella desarrollados recientemente, se han obtenido grandes avances en la caracterización y tratamiento de diversas enfermedades. No obstante, la comprensión de como los distintos niveles moleculares se relacionan para llevar a cabo todas las funciones vitales sigue siendo el mayor reto de las ciencias de la vida (*e.g.* se calcula que sólo se conoce un 10 % de todas las interacciones entre proteínas en humanos [1]).

Desde que se publicase la secuencia completa del genoma humano, allá por el año 2000, la cantidad de datos biológicos ha ido en aumento de una forma exponencial y el coste de las secuenciaciones de genomas se ha reducido también a un ritmo vertiginoso (véase la figura 1.1), hasta llegar a convertir a la Biología en una de las ramas más productivas de la Ciencia, dando lugar a la llamada 'era postgenómica'. Por otro lado, cabe destacar que toda esta información se encuentra almacenada (y normalmente disponible *online*) en forma de series casi interminables de: nucleótidos (con información respecto a las secuencias de ADN), aminoácidos (representando secuencias de proteínas), *arrays* de expresión génica, ontologías, datos de estructuras moleculares, redes de interacción de proteínas, etc.

Todos estos datos esconden información muy útil para la comprensión de los seres vivos, así como para el estudio e identificación de nuevas terapias con las que abordar diferentes tipos de enfermedades. Es entonces cuando se requiere de una unión entre las técnicas más avanzadas de computación (supercomputación, computación paralela, minería de datos, análisis automático, inteligencia artificial, etc.) y los conocimientos biológicos para poder interpretar de manera útil tal cantidad de datos almacenados. Existen diversas ramas dentro del campo conocido como 'Bioin-

[1]La transcriptómica, la interactómica, y la metabolómica forman parte, entre otras, de las llamadas ciencias 'ómicas': neologismo del inglés (*omics*) que se utiliza para referirse al estudio de la totalidad; en estos casos a la totalidad del ARN (ARNr, ARNt, ARNm, ARNi, miARN), la totalidad de las interacciones biomoleculares y la totalidad de los procesos químicos que involucran metabolitos, respectivamente.

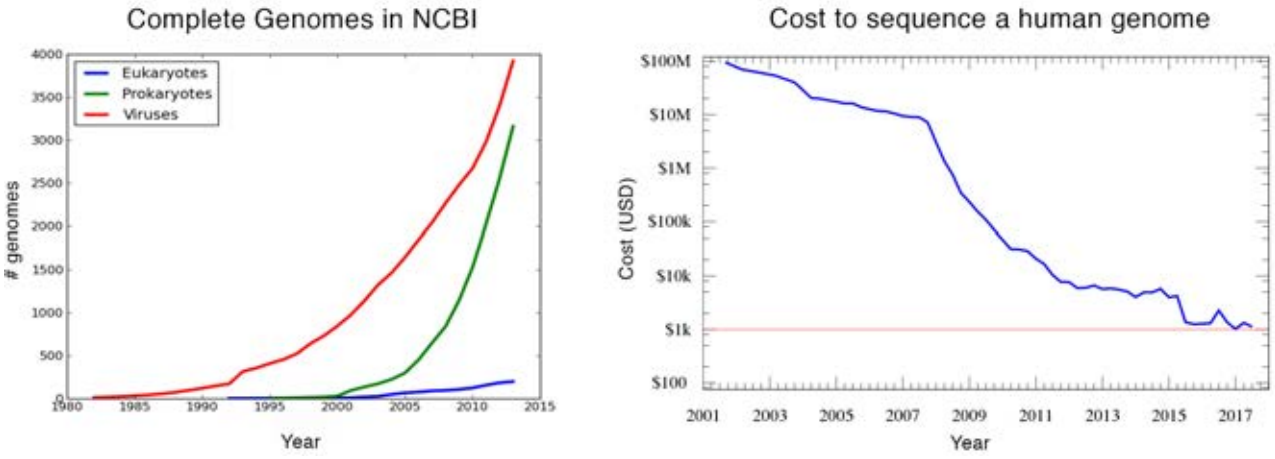


Figura 1.1: **Crecimiento de datos genómicos y reducción de costes de secuenciación.** Izquierda: evolución del número de genomas secuenciados en NCBI (fuente NCBI); derecha: evolución del coste de secuenciar un genoma humano completo (fuente *National Human Genome Research Institute*).

formática', que básicamente podemos definir como: la disciplina que aplica herramientas computacionales al almacenamiento, gestión y estudio de datos biológicos. Se trata de un campo relativamente nuevo, cuyas áreas más destacables son: 1) la **Genómica computacional**, la cual se centra en estudiar genomas (conjunto de todo el material genético de un organismo), por ejemplo para la búsqueda de nuevos genes implicados en procesos biológicos (o enfermedades) o realizar comparativas de genomas de diferentes especies. Recordemos que un gen es una unidad de información del ADN que codifica un producto funcional, *e.g.* una proteína; 2) la **Bioinformática estructural**, que focaliza sus esfuerzos en el estudio de estructuras moleculares, por ejemplo mediante la predicción de la morfología tridimensional de proteínas a partir de su secuencia de aminoácidos o buscando moléculas útiles para ser usadas como fármacos en el tratamiento de determinadas enfermedades, en base a sus estructuras y puntos de unión; y 3) la **Biología de Sistemas** que, pese a ser un campo bastante amplio, se podría definir como el estudio de sistemas biológicos complejos (como por ejemplo el metabolismo o las rutas de señalización celular) de manera holística, su modelización matemática teniendo en cuenta los elementos internos y externos del sistema -así como sus interconexiones- y la realización de predicciones basadas en estos modelos que lleven al

descubrimiento de nuevos elementos funcionales o propiedades globales no presentes en las partes (llamadas **propiedades emergentes**).

El presente trabajo se desarrolla fundamentalmente dentro del área de la **Biología de Sistemas**, modelando diferentes tipos de sistemas biológicos complejos en redes de asociación y analizando sus propiedades para la realización de predicciones de nuevas relaciones funcionales. Principalmente se han estudiado:

- 1) Relaciones filogenéticas.
- 2) Redes de interacción de proteínas.
- 3) Redes biomédicas de asociaciones entre: mutaciones, pacientes y ontologías médicas (fenotipos).

Al tratarse este de un trabajo multidisciplinar no está de más recordar de manera sucinta el conocido como **dogma central de la Biología**, pues algunas de sus áreas son objeto de las investigaciones aquí expuestas: los genes formando parte del ADN en la célula (**nivel genómico**) sufren un proceso de transcripción, que implica que su información es copiada en una cadena de ARN (**nivel transcriptómico**) necesaria para producir las proteínas que codifican (**nivel proteómico**), en un proceso conocido como traducción (véase la figura 1.2).

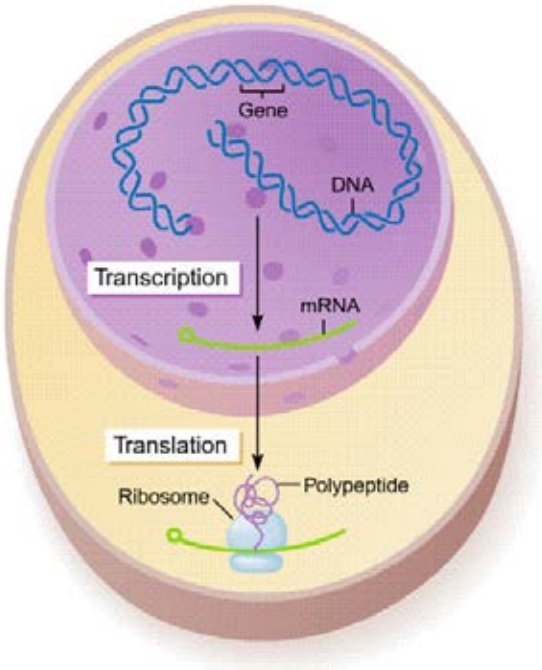


Figura 1.2: **Dogma central de la Biología.** Proceso de transcripción del ADN en ARNm y de traducción de la información genética en proteínas. En este caso se muestra el ejemplo de una célula eucariota, donde la transcripción tiene lugar en el núcleo celular y la traducción en el citoplasma.

1.1. Biología Molecular de Sistemas

Es obvio el hecho de que la vida, en todas sus expresiones y niveles, es producto de un gran número de interrelaciones complejas. La mayoría de las funciones de los sistemas biológicos no se pueden deducir de una simple suma de sus componentes; pero para llegar a la comprensión de este hecho la Ciencia ha pasado por diferentes fases en cuanto a su forma de abordar el problema. En este sentido, en el siglo XVII, con la aparición del pensamiento cartesiano suscitado por el famoso filósofo, físico y matemático francés René Descartes [68], se establecen las bases de un método protocientífico, según el cual los problemas complejos pueden desglosarse en un conjunto de problemas más simples cuyo análisis permite posteriormente llegar a comprender el todo del que forman parte, mediante un procedimiento racional de suma e integración lineal de dichos componentes. Esta visión es la base del 'reduccionismo'.

En Biología, la corriente reduccionista (y mecanicista) alcanzó su hito como método aplicado al estudio de la vida en 1912 con la publicación del libro *La Concepción Mecánica de la Vida* de Jacque Loeb [69], donde se consideraba que los organismos no eran más que máquinas complejas. Ya Aristóteles, en su libro *VIII* de Metafísica, se oponía a esta visión, alegando que 'el todo está siempre por encima de sus partes y es más que la suma de todas ellas'. Pero aunque algunos científicos apoyaban tal principio en las funciones de los sistemas vivos, esta línea de pensamiento -denominada 'holística'- no consiguió despegar como una corriente conceptual y metodológica de peso. Hay que esperar hasta 1926 para encontrar la primera referencia importante al concepto 'holismo' en Biología [70]; según el cual las propiedades de los sistemas no pueden reducirse a las de sus partes constituyentes[2].

[2]Parte de la información sobre la historia de la Biología Molecular de Sistemas ha sido extraída y sintetizada del proyecto docente presentado por el Dr. D. Juan Antonio García Ranea al concurso para el acceso a plaza de profesor titular en la Universidad de Málaga (BOE num. 230 de 23 de Septiembre de 2016 - trabajo no publicado). Del mismo modo se ha extraído información de las siguientes publicaciones: 1) Medina MÁ. *Systems biology for molecular life sciences and its impact in biomedicine*. Cell. Mol. Life Sci. 2013; 70: 1035-1053. 2) Encuentros en la Biología Vol.IV. no.136. Diciembre 2011. 3) Proyecto docente del Dr. D. Miguel Ángel Medina Torres.

1.1. Biología Molecular de Sistemas

La *Teoría General de Sistemas* [71], desarrollada por el biólogo Ludwig von Bertalanffy entre los años 40 y 60, postula que un sistema vivo se caracteriza por las interacciones de sus componentes y la no linealidad de esas interacciones. Comenzó por entonces la modelización matemática de sistemas biológicos y su análisis, como un leve despertar de lo que posteriormente sería la Biología de Sistemas, aunque estas técnicas parecían únicamente despertar el interés de científicos de perfil matemático.

En el libro de 1968 *La estructura irreductible de la vida* [72], Michale Polanyi sostiene que los sistemas vivos tienen una estructura jerárquica, y esta se mantiene a través de complejas redes de interacciones entre los elementos dentro de cada nivel y entre los diferentes niveles. Polanyi critica la visión reduccionista y mecanicista del mundo que tenía la Ciencia y defiende que las funciones vitales son **propiedades emergentes** de un sistema complejo como es el organismo. En cualquier caso, visiones reduccionistas y holísticas no son excluyentes, sino que pueden ser complementarias. Esta síntesis entre reduccionismo y holismo fue formulada por el eminente científico François Jacob (premio Nobel de Fisiología o Medicina en 1965) en su libro *La lógica de lo viviente* (1974) [73].

Tras el cambio de paradigma teórico, el principal problema era poder pasar de una disciplina con modelos puramente matemáticos a otra con mayor protagonismo práctico en Biología, y más concretamente en lo que concierne a este trabajo, en Biología Molecular. Este paso requería de la aparición de nuevas tecnologías que permitieran ensayos experimentales sistémicos (holísticos): la obtención masiva de datos biológicos y su posterior análisis. Durante las décadas de los 70 y los 80 se comienzan a desarrollar las técnicas que darán soporte posteriormente a los ensayos de alto rendimiento en la transcriptómica y la proteómica [74–76]. Y no es hasta la década de los 90 cuando la tecnología se encuentra lo suficientemente avanzada como para plantearse la secuenciación completa de genomas. El primer eucariota completamente secuenciado fue la levadura *S. cerevisiae*, en un trabajo publicado en 1996: *Life with 6000 genes*, por Goffeau *et al.* [77]. Ya a finales del milenio, en torno a 1999, encontramos las primeras publicaciones con raíces en la Biología Molecular de Sistemas tal y como la conocemos (*e.g.* enfocadas a la comparación de perfiles filogenéticos de

manera sistémica [78]). Aparecen las primeras revistas del sector y se acuña el término **Biología de Sistemas** como tal.

El cambio de paradigma metodológico viene, efectivamente, produciéndose desde principios de siglo (coincidiendo con la secuenciación del genoma humano y los primeros intentos de abordar su complejo análisis) hasta nuestros días en forma de revolución científica. En 2004, en el trabajo *Complexity in biology. Exceeding the limits of reductionism and determinism using complexity theory*, Fulvio Mazzocchi expone que la metodología reduccionista es incapaz de trabajar con la enorme cantidad de datos producida por las ciencias '-omicas' y la complejidad de los sistemas que estas revelan, y que el tiempo del método de Descartes pasó y es imprescindible en las ciencias de la vida evolucionar hacia modelos y herramientas diferentes [79].

Podemos decir que la **Biología Molecular de Sistemas** (el área de la Biología de Sistemas dedicada a los estudios moleculares) fundamenta sus métodos en la abstracción racional de hechos experimentales (*e.g.* las interacciones entre proteínas). Estas aproximaciones basadas en modelos teóricos se sitúan en un escenario sistémico pero a su vez reduccionista (*e.g.* representando a una proteína como un nodo de una red y a una interacción entre dos de ellas como una arista). A partir de ahí se aplican modelos deductivos para tratar de poner a prueba la hipótesis previa, lo cual llevará a unas conclusiones que de nuevo deberán ser comprobadas experimentalmente. Nos encontramos ante los primeros pasos de esta joven ciencia y aún se arrastran modelos de experimentación, tecnologías y análisis de paradigmas anteriores.

Quizás, la mayor limitación de las metodologías sistémicas, hoy en día, se encuentre más en la visión de la comunidad científica sobre cómo afrontar las investigaciones que en las limitaciones técnicas.

Tal como se ha mencionado, la Biología de Sistemas (y dentro de ésta la Biología Molecular de Sistemas) estudia los sistemas biológicos complejos, a diferentes niveles. Los **sistemas complejos** son aquellos compuestos por varias partes enlazadas e interconectadas. Sistemas que han de ser analizados de una forma diferente, con un enfoque holístico, difícil de alcanzar hasta hace relativa-

mente poco tiempo por la falta de medios computacionales. Sistemas con propiedades emergentes y en los cuales la comprensión del funcionamiento de las interacciones entre elementos es tanto o más importante que la de los componentes en sí mismos.

Como ejemplo de **sistema complejo** que requiere un enfoque sistémico propio de esta disciplina está el de las redes de regulación génicas, donde no solamente hay genes que se expresan produciendo ARN, el cual a su vez produce proteínas (véase la figura 1.2), sino que además también existen genes que regulan la transcripción de otros genes, genes que se auto-regulan, densidades de transcrito o de proteína que activan otros genes, etc. Y todo ello ocurriendo simultáneamente.

Los métodos, relativos a la Biología Molecular de Sistemas, aplicados en este trabajo parten de un conjunto de datos que modelan un sistema biológico complejo (*e.g.* en forma de red) y llevan a cabo los análisis pertinentes sobre el mismo que permiten la predicción funcional de nuevos componentes. Utilizando 'el todo' para obtener información de 'la parte'. En este tipo de análisis sistémicos de los sistemas complejos surgen, por la propia naturaleza de las relaciones entre los elementos, lo que se conoce como **propiedades emergentes**: aquellas que sólo existen por la interacción de los componentes del sistema y no se explican únicamente por la suma de dichos componentes.

Como ejemplo de propiedad emergente de un sistema complejo se puede tomar la de los diferentes significados (propiedades) que adquiere una oración en un determinado lenguaje (sistema complejo) según la relación que exista entre las palabras que la componen (y no dependiendo exclusivamente de los componentes -las palabras- en sí mismos). Un claro ejemplo lo compondrían estas dos frases: i) 'Me baño en el río' y ii) 'Me río en el baño'. La diferencia en significados emerge de las relaciones entre las palabras, las cuales son exactamente las mismas en ambos escenarios.

1.1.1. Aplicación de la Biología Molecular de Sistemas en el presente trabajo

Este trabajo se enmarca dentro de la **Biología Molecular de Sistemas** y hace uso de sus datos, técnicas y herramientas. Entre los datos utilizados podemos destacar los relativos a los **árboles filogenéticos** (véase el capítulo 1.2), las **redes de interacción de proteínas** (véase el capítulo 1.3); así como **datos sobre pacientes con enfermedades raras** (véase el capítulo 1.4), tales como: **información sobre mutaciones**, obtenidas de bases de datos específicas (DECIPHER [62]) o **datos ontológicos de fenotipos patológicos** (haciendo uso de HPO[52]).

Del mismo modo que los seres vivos que pueblan un ecosistema: animales, plantas, microorganismos, junto con sus hábitats y formas de vida constituyen un sistema complejo con unas propiedades que no poseen los organismos individuales por sí mismos (véase la figura 1.3), un cuerpo humano (o de cualquier otro ser vivo) y su funcionamiento son mucho más que sus componentes aislados. Es aquí donde aparece el nuevo paradigma científico en el que se enmarca este trabajo: los sistemas complejos y la Ciencia de Sistemas: Biología de Sistemas [80, 81], Genética de Sistemas [82], Medicina de Sistemas [83–85], etc.

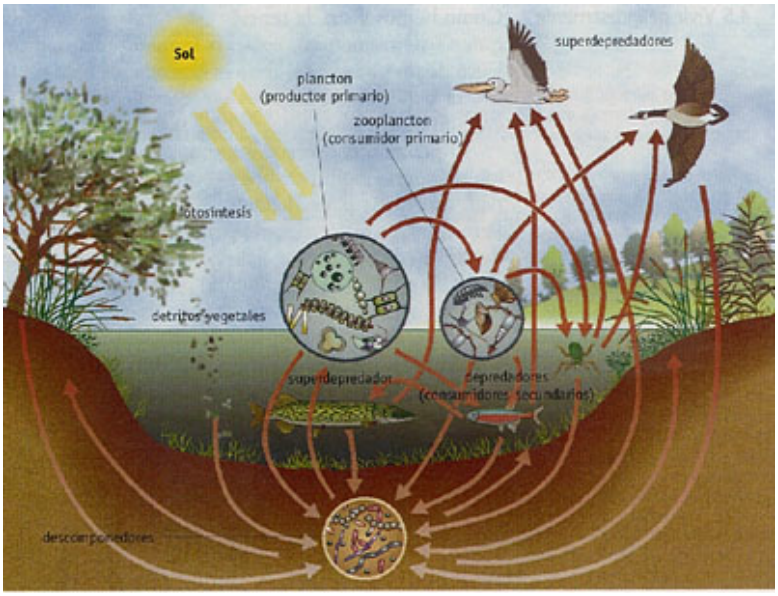


Figura 1.3: **Partes de un ecosistema (sistema complejo).**

El prestigioso científico en redes biológicas Albert-László Barabási destaca el hecho de que la Medicina de Sistemas se está convirtiendo en un enfoque esencial para establecer relaciones moleculares entre diferentes fenotipos patológicos, identificar nuevos genes en enfermedades y encontrar el efecto de mutaciones y su asociación con enfermedades [3]. La Medicina de Sistemas hace uso de la teoría de redes y ha podido determinar algunas propiedades generales inherentes a las redes biológicas [4, 5].

La aplicación de las técnicas de Biología Molecular de Sistemas en este trabajo, orientadas a la obtención de conocimiento dentro del área de la biomedicina, se ha llevado a cabo mediante el análisis de redes de interacción entre entidades, en diferentes ámbitos de aplicación: desde las redes de interacción de proteínas y sus relaciones filogenéticas, pasando por los pacientes y la red que conforman, hasta llegar al área de la medicina diagnóstica, relacionando fenotipos clínicos con genotipos en diferentes síndromes. La figura 1.4 muestra los citados ámbitos de aplicación de las técnicas de la Biología Molecular de Sistemas en este trabajo.

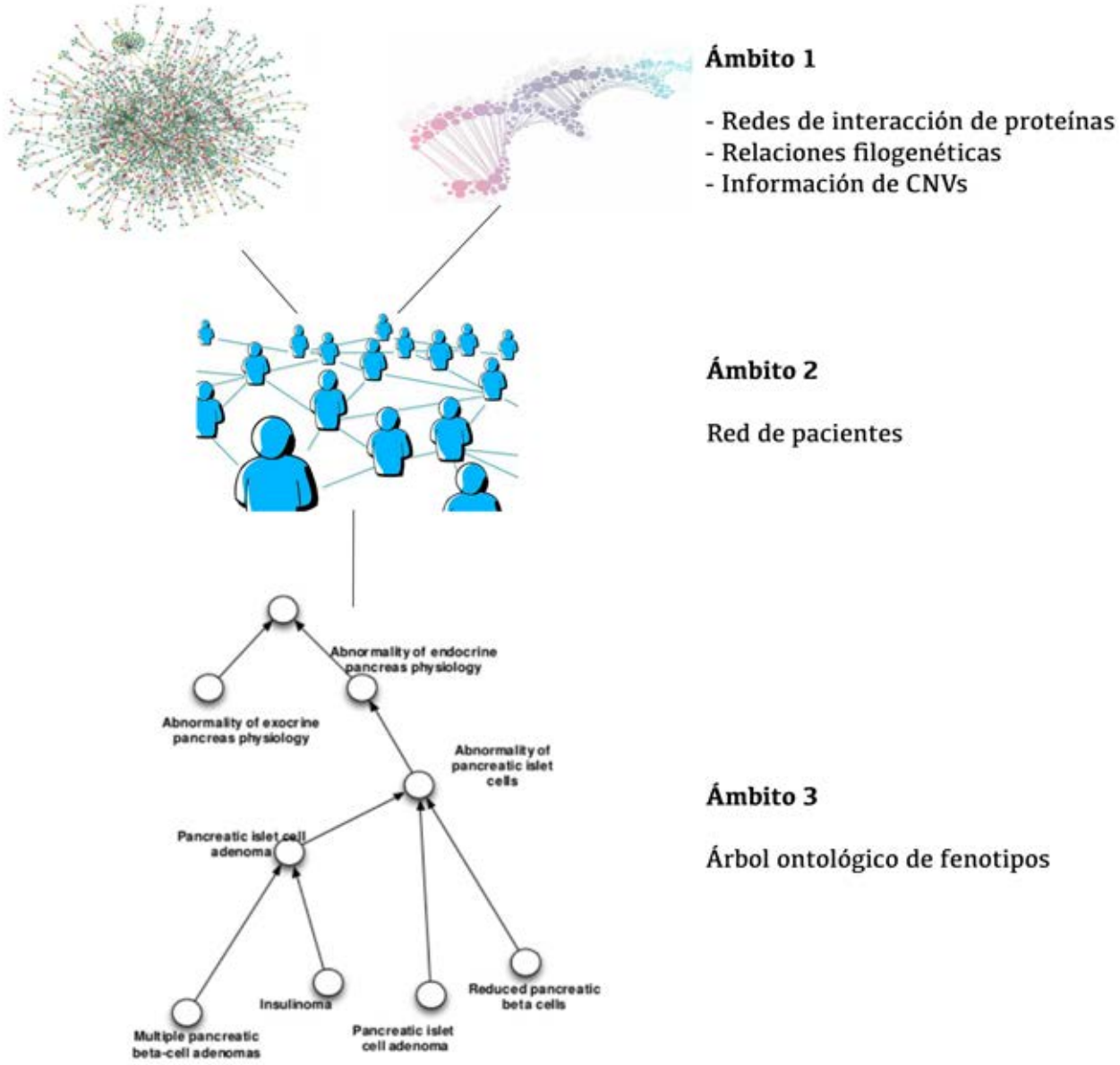


Figura 1.4: **Ámbitos en los que se ha aplicado la Biología Molecular de Sistemas en esta Tesis Doctoral:** 1) Ámbito molecular: en redes de interacción de proteínas, relaciones filogenéticas y estudio de mutaciones; 2) Ámbito de red de pacientes; 3) Ámbito de relaciones fenotípicas: árbol ontológico y estudio de características clínicas.

1.2. Árboles filogenéticos

En el contexto de los estudios aquí presentados los árboles filogenéticos fueron utilizados como forma de catalogar y medir las distancias evolutivas (a nivel molecular) en la familia de proteínas RAS y su posterior comparación con las relaciones funcionales en el interactoma humano (véase el capítulo 5).

La idea de la construcción de un *árbol de la vida* es bastante antigua. En sus inicios consistía simplemente en la representación de una progresión desde las formas de vida más primitivas hasta las más complejas. Las primeras representaciones de ramas filogenéticas procedían de datos paleontológicos, simbolizando relaciones entre animales y plantas y fueron desarrolladas por Edward Hitchcock en su libro *Elementary Geology* en 1840.

Fue Charles Darwin, en 1859, el que realizó una de las primeras ilustraciones en forma de árbol y popularizó el término 'árbol evolutivo' en su publicación *El Origen de las Especies*. Un siglo y medio después los biólogos evolucionistas siguen haciendo uso de diagramas en forma de árbol para expresar la evolución, ya que es la mejor forma para representar todos los datos necesarios, así como los conceptos de especiación que ocurren durante los procesos adaptativos.

La filogenética es la ciencia que estudia la historia evolutiva de los organismos, más concretamente de los genes de dichos organismos. A nivel molecular, la construcción de árboles filogenéticos permite plasmar visual y numéricamente diferencias graduales entre moléculas de un mismo organismo (parálogas) o de distintos organismos (ortólogas). Con esto se pueden establecer grados de semejanza en grupos de moléculas (bien sean ADN o proteínas). A partir de sus secuencias se calculan matemáticamente las similitudes y diferencias y, junto con los datos evolutivos conocidos, se puede reconstruir esta especie de 'árbol evolutivo molecular' y extraer información del mismo.

A lo largo de la historia evolutiva, en un genoma, se producen procesos de *duplicación génica*. Este evento ocurre cuando, derivado de un fallo de replicación a la hora de copiar el genoma en una división celular, se produce un escenario en el cual se ha copiado por dos veces el mismo gen.

Tras producirse esta duplicación (que puede ser total o parcial, si implica al gen entero o únicamente a un fragmento del mismo), lo que ocurre es que coexisten 2 fragmentos del genoma que puede que codifiquen la misma proteína. Dependiendo de las mutaciones que se sucedan en el gen original y en el gen duplicado, las secuencias cambiarán, en un proceso denominado *divergencia* (y del mismo modo la proteína que codifiquen), dando lugar a proteínas con un origen común pero con una secuencia y funcionalidad diferentes. En la figura 1.5 se muestran ejemplos de los posibles escenarios tras una duplicación génica. Un análisis de las secuencias de ambas proteínas (o de las secuencias de ambos genes) proporciona una visión de las divergencias entre ambos y permite cuantificar su grado de *lejanía* filogenética. Si esto se aplica a familias enteras de proteínas (o genes) se puede reconstruir, en forma de árbol, la mencionada historia evolutiva -a nivel molecular- entre ellas.

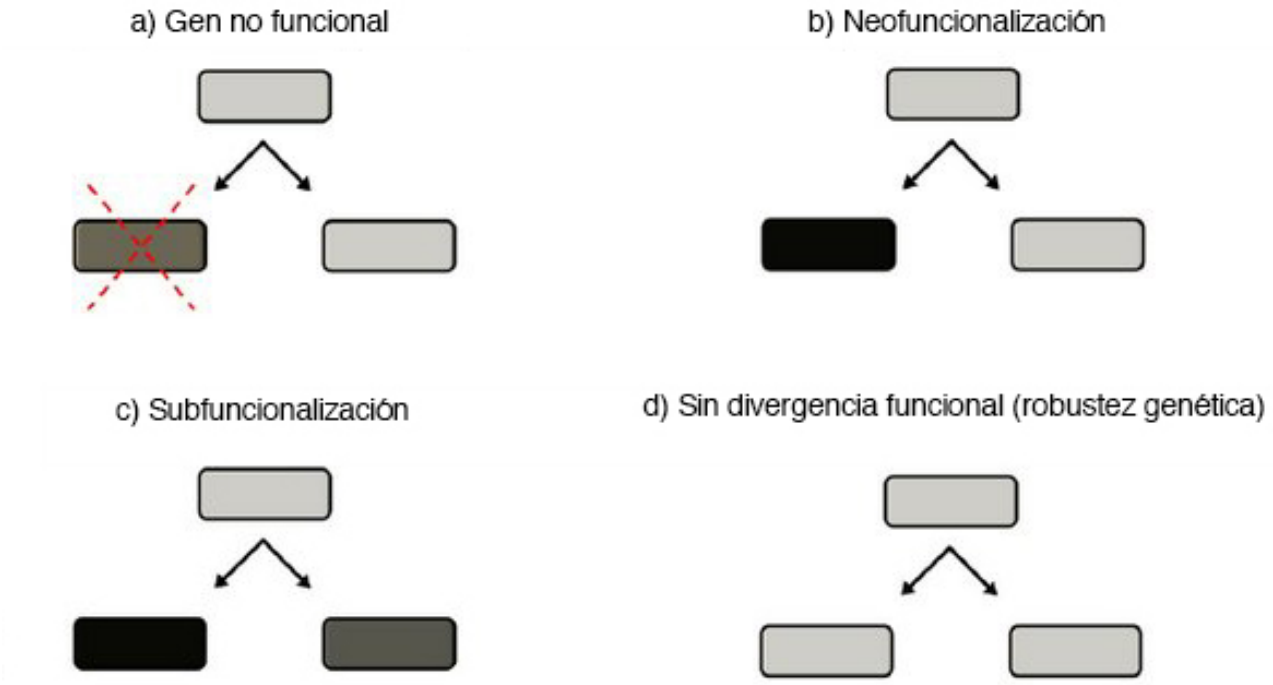


Figura 1.5: **Escenarios tras una duplicación génica**: **a)** nuevo gen no funcional. **b)** una copia del gen duplicado retiene la función original mientras que la otra adquiere una nueva función: este escenario es el que hace mayoritariamente de motor del cambio evolutivo. **c)** ocurren mutaciones en ambas copias del gen duplicado, adquiriendo funciones complementarias, repartiendo las funciones del gen ancestral. **d)** nuevo gen con la misma funcionalidad que el gen original: simplemente se trata de una redundancia que dota de robustez, ante mutaciones en una de las copias, a la expresión de la proteína codificada en el gen. Adaptado de Conrad y Antonarakis (2007).

Hoy en día, con las nuevas técnicas bioinformáticas, la cantidad de datos filogenéticos que se genera y almacena en bases de datos *online* es muy grande. De hecho, existen estudios que incluyen la reconstrucción de filomas de especies (conjunto de todos los árboles filogenéticos). En este sentido, Gabaldón *et al.* hicieron pública en 2007 la base de datos PhylomeDB [86, 87], la cual contiene el filoma humano (incluyendo ortólogos de 39 especies).

Se pueden encontrar diversas fuentes de datos disponibles para la obtención de información filogenética; de entre ellas las más conocidas son: TreeBASE [88] y TreeFam [89].

En la parte visual de estas bases de datos, se suele representar la información de los cambios genéticos (diferencias en las secuencias) que permiten inferir relaciones evolutivas en forma de árbol (véase la figura 1.6) donde cada ramificación significa un cambio evolutivo. Para enriquecer los datos, cada rama suele estar acompañada de un valor (peso) que indica la distancia entre un punto del árbol filogenético y su ancestro más cercano, ya que un cambio evolutivo puede ser leve o severo. Estos árboles filogenéticos se construyen a partir de secuencias biológicas (*e.g.* los aminoácidos que forman las proteínas) alineadas de manera múltiple, para las cuales se calculan matrices de distancias entre ellas utilizando métodos tales como *Neighbor-joining* (algoritmo bioinformático de tipo '*bottom-up*' basado en clusterización).

A nivel práctico, se pueden encontrar los árboles en las bases de datos en forma de ficheros informáticos codificados siguiendo un formato establecido que permite representar el árbol en texto plano. Este formato (el más utilizado de ellos) es el formato *Newick Standard* [90], establecido en 1857, y cuya correlación con la estructura del árbol se puede observar en la figura 1.7.

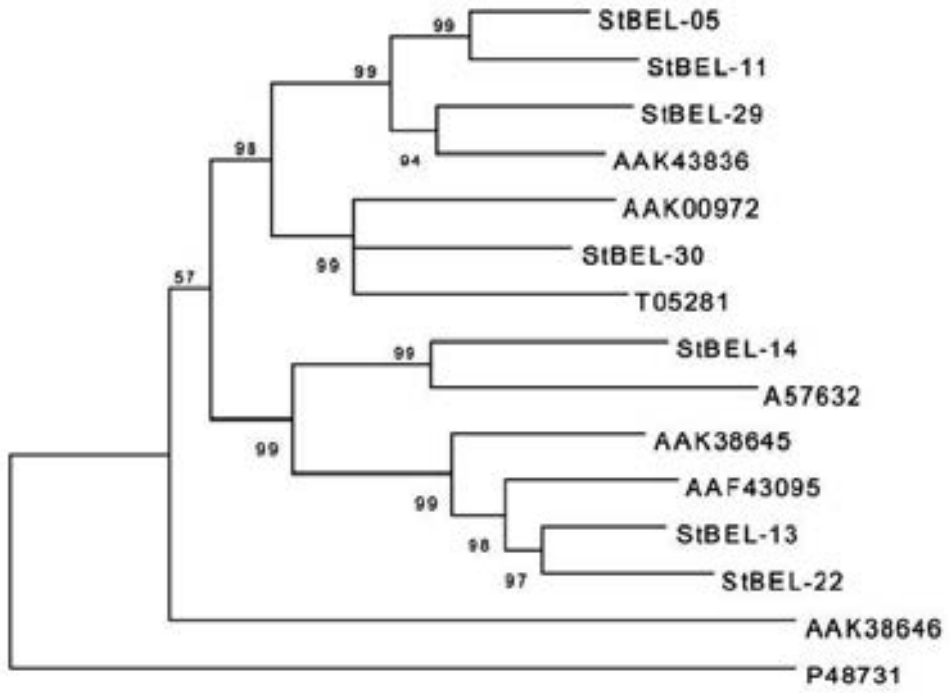


Figura 1.6: **Ejemplo de árbol filogenético con sus ramas ponderadas.** En este caso se trata de la representación para la familia de proteínas BEL-1, en la patata.

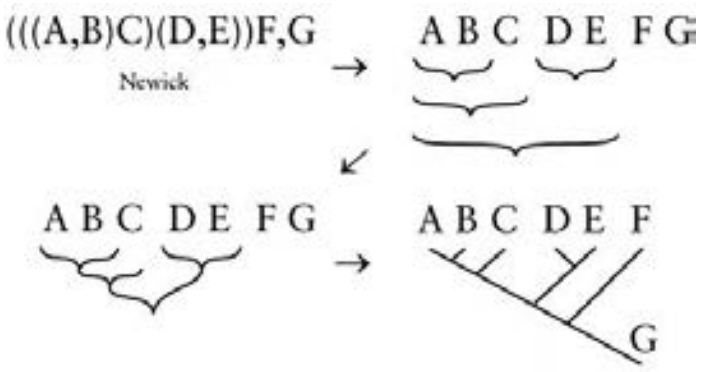


Figura 1.7: **Ejemplo de codificación en formato *Newick Standard* para árboles filogenéticos.**

1.3. Redes de interacción de proteínas

En el presente trabajo las interacciones conocidas entre proteínas en *Homo sapiens* han sido representadas mediante redes (siendo los nodos las propias proteínas y las aristas las interacciones entre ellas) y a dichas redes se les han aplicado algoritmos de análisis con la finalidad de medir las *distancias* entre pares y poder utilizar dicha información en dos de los estudios presentados en esta Tesis Doctoral: i) construyendo predictores para la identificación de nuevas proteínas implicadas en procesos moleculares (véase el capítulo 4) y ii) comparando las diferencias evolutivas con las funcionales dentro de una familia de proteínas (véase el capítulo 5).

Gran cantidad de sistemas pueden ser representados en forma de red. Se trata de un tipo de modelado muy útil. Y en el caso concreto de la Biología el número de sistemas que actualmente son abstraídos de esta forma es muy elevado: desde rutas metabólicas (representaciones de los substratos metabólicos y los productos, unidos mediante aristas dirigidas), hasta las redes de regulación génica (incluyendo genes activadores e inhibidores, transcritos y proteínas), pasando por las redes de interacción de proteínas, entre otras. Estas últimas se detallan a continuación.

Las redes de interacciones entre proteínas (PPI) tratan de reconstruir, a partir de diferentes fuentes de datos disponibles, todas las interacciones que se producen en los organismos vivos entre proteínas, *i.e.* el interactoma [91], representando a las proteínas como nodos en la red y a sus interacciones como aristas (la figura 1.8 muestra el aspecto de una de estas redes). El estudio de las redes de interacción de proteínas juega un papel cada vez más importante en la comprensión de los mecanismos celulares y las enfermedades. Los primeros análisis de interacciones para la reconstrucción de rutas metabólicas, tal como los conocemos hoy en día, datan de la década de los cuarenta, pero fue a finales de los años noventa cuando la informática impulsó un crecimiento exponencial en este tipo de análisis. En la actualidad existe una importante cantidad de bases de datos disponibles con esta clase de información y cada vez son más las técnicas que se aplican a gran escala para analizar las ingentes cantidades de datos que albergan.

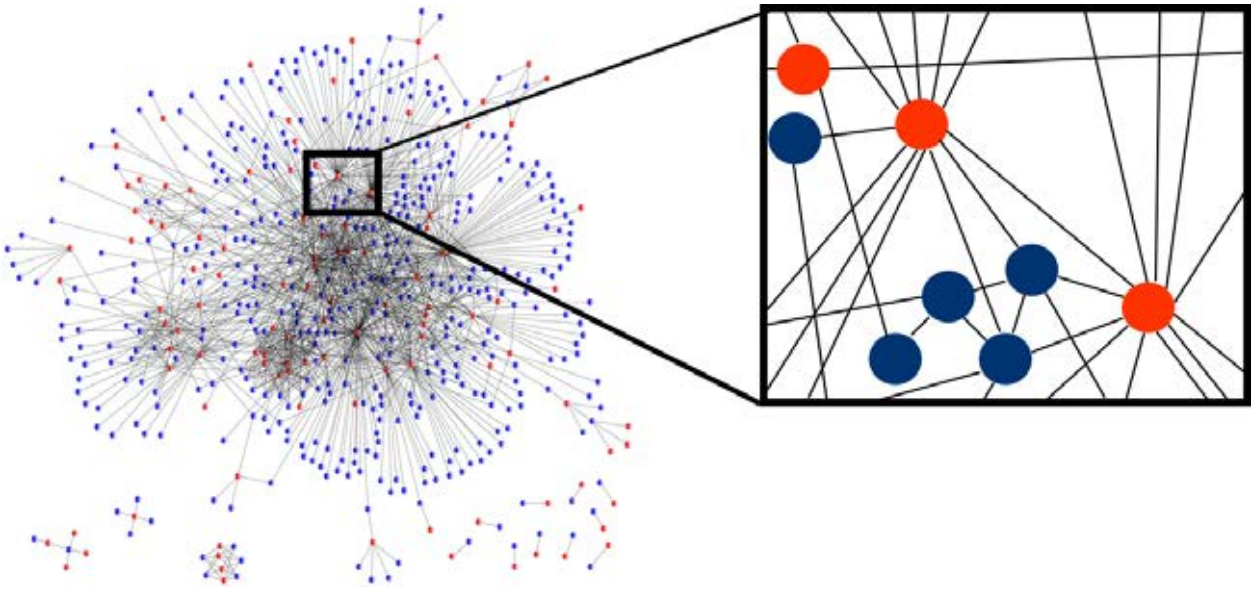


Figura 1.8: **Ejemplo de red de interacción de proteínas**: cada nodo representa una proteína y cada arista una interacción entre ellas.

Uno de los principales retos de la Biología Molecular de Sistemas es la reconstrucción de todas las interacciones entre proteínas que se producen en las células (el llamado interactoma). En humanos se estima que sólo se ha caracterizado experimentalmente un 10 % actualmente [1, 92–94]. Avances en este sentido ayudarían a comprender la maquinaria interna de la célula, permitiendo el diseño de fármacos o la caracterización de los mecanismos asociados a enfermedades celulares.

Las interacciones físicas entre proteínas son identificadas usando diferentes técnicas (*e.g.* coinmunoprecipitación, ensayo de doble híbrido, inmunoprecipitación cuantitativa); no obstante, existen otras fuentes de generación de redes teóricas de interacciones proteicas. Mediante diferentes técnicas de computación: análisis automático de coevolución, métodos basados en co-ocurrencia de perfiles, coincidencia de dominios estructurales, análisis de vecindad genómica, datos filogenéticos, predicciones basadas en secuencias, análisis a través de minería de textos científicos, métodos basados en homología, similitud semántica, etc.; se pueden generar nuevas bases de datos de interacciones con cierto grado de certeza asociado [95].

Cabe destacar que la construcción de los modelos de red no es infalible, está sujeta a errores derivados de la propia experimentación. Existen, por ejemplo, interacciones que son comprobadas *in vitro* y dadas por buenas cuando realmente no ocurren *in vivo* debido a que no se da la colocalización necesaria de las proteínas en la célula o a otros factores no controlables en el laboratorio. En ocasiones, también se ha llegado a resultados contradictorios de interacción para grupos de proteínas. Es por todo lo anterior que a la hora de analizar sistémicamente estas bases de datos con interacciones entre proteínas es recomendable integrar más de una fuente, para dotar de mayor consistencia a los resultados. La combinación de múltiples fuentes de evidencia independientes proporciona un grado de fiabilidad mayor a los modelos de redes de interacción así obtenidos.

En este trabajo se analizaron diversas bases de datos y se hizo una selección para su estudio, basada en: la fiabilidad, el origen de la información y la cobertura con los datos problema en cada caso. Se generaron los modelos de red correspondientes y se implementaron métodos de análisis para su explotación.

1.3.1. Distancias en redes

Un grafo (véase la figura 1.8) es una abstracción matemática de una red y viene definido por: i) un conjunto de nodos y ii) un conjunto de aristas. Dicha abstracción es capaz de representar complejas estructuras y es una pieza clave, actualmente, en la Bioinformática y en la Biología de Sistemas. El estudio matemático de redes es el objeto de la **Teoría de grafos**, la cual es uno de los pilares fundamentales de la **Matemática discreta**.

La teoría de grafos ha ido ganando peso dentro de las matemáticas aplicadas en los últimos años, y recientemente, junto con los avances tecnológicos en los sistemas computacionales, ha sufrido una revolución en sus formas de aplicación. La capacidad actual para generar grandes redes (entre ellas las biológicas) y los nuevos supercomputadores han producido un cambio, pasando de sencillos análisis conceptuales de pequeños grafos a macroestructuras que han de ser, necesaria-

mente, analizadas de manera sistémica y global, dada la imposibilidad de entrar al detalle topológico, mediante fórmulas estadísticas que muestren tendencias generales. Toda la metodología que actualmente se desarrolla en este ámbito viene a sustituir a los análisis visuales sobre las sencillas redes que se llevaban a cabo en épocas anteriores, pero con un potencial infinitamente más elevado: las matemáticas permiten llegar a conclusiones sobre 'cómo es una red' de enorme tamaño sin la capacidad de poder 'verla'.

Para explotar las redes y extraer información sobre las relaciones de sus elementos, se requiere de un método de análisis que permita determinar la *distancia* dentro de la red entre cada par de nodos (si se trabaja con redes de interacción de proteínas se podría hablar de la distancia funcional entre dos de ellas). Este análisis de la red, a nivel computacional, se puede realizar de diversas formas: a) mediante algoritmos de búsqueda de los caminos más cortos entre nodos en la red, b) mediante algoritmos probabilísticos de difusión, o c) mediante comparaciones de perfiles; entre otros.

En el capítulo 3.2.1, de Materiales y métodos generales, se ofrece una visión panorámica de los tipos de algoritmos que existen para medir distancias en redes, así como de sus principales características, estableciendo de este modo una base conceptual de la metodología y una justificación teórica de la selección del conjunto de métricas que finalmente se aplican en los estudios incluidos en esta Tesis Doctoral.

1.3.2. Predictores funcionales basados en análisis de redes de interacción de proteínas

Uno de los objetivos que se persigue dentro del campo de estudio de las interacciones entre proteínas es el reconocimiento de nuevas interacciones que estén implicadas en la aparición de enfermedades o en su desarrollo. La inmensa mayoría de las interacciones entre proteínas continúan siendo un misterio para la comunidad científica, como ya se ha destacado anteriormente; y a pesar

de los esfuerzos que se han llevado a cabo en las últimas décadas, se estima que únicamente se conocen experimentalmente en torno al 10 % de las interacciones proteína-proteína en *Homo sapiens*, y además, de un tercio de las proteínas humanas se desconoce todas sus interacciones, y más de la mitad de los genes humanos conocidos que codifican proteínas no poseen registros experimentales que evidencien su funcionalidad [1, 94]. El análisis de estas funcionalidades e interacciones desconocidas se ha visto potenciado con la aparición de las técnicas de alto rendimiento, pero desafortunadamente la complejidad de los ensayos necesarios y los costes de los mismos hacen que estos sean inasumibles para la mayoría de grupos de investigación. Además, la búsqueda de nuevos posibles casos de estudio implica amplios tanteos experimentales y arduas búsquedas bibliográficas.

Las limitaciones comentadas anteriormente hacen que exista un sesgo de exceso de investigación en un conjunto de proteínas concretas que son ya bien conocidas [46]. Tendencia que se extiende a la investigación sobre fármacos y terapias, la cual también se concentra en un grupo reducido de dominos de proteínas [96].

Es por ello que se torna importante la predicción funcional de genes/proteínas a la hora de abordar una investigación, y para ello, la aplicación de predictores computacionales capaces de priorizar un subconjunto de genes o proteínas, con cierta fiabilidad, sobre los cuales llevar a cabo la experimentación que valide dichas predicciones. Es esta una herramienta de incalculable valor, que reduce tiempos y costes de manera notable. En este contexto, se ha desarrollado en la presente Tesis Doctoral un sistema de predicción funcional basado en el modelado y explotación de redes de interacción de proteínas [97], cuya metodología (para su implementación y validación) se detalla en el capítulo 3.3 de Materiales y métodos generales. El estudio asociado a este predictor corresponde al capítulo 4 de este manuscrito.

1.4. Identificación de relaciones genotipo-fenotipo en enfermedades raras

En el capítulo 6 se detalla otro de los estudios que avalan esta Tesis Doctoral, en el cual se analizaron datos de pacientes con enfermedades raras con la finalidad de detectar relaciones genotipo-fenotipo que pudieran ayudar en el diagnóstico clínico y la caracterización de este tipo de patologías.

1.4.1. Las enfermedades raras

Una enfermedad rara es aquella que afecta a una pequeña proporción de la población. La definición cuantitativa varía dependiendo del organismo que la establezca, *e.g.*: en Estados Unidos se define como la que padece un número de personas menor a 200.000 a nivel global [98], en Japón como la que afecta a menos de 50.000 y en la Unión Europea como aquella que padece 1 de cada 2.000 personas [6]. En términos porcentuales, se corresponden, más o menos, con tasas que van desde el 0,01 % al 0,05 % de la población. El hecho singular de su baja frecuencia las hace muy difíciles de diagnosticar y tratar, así como de desarrollar la investigación clínica asociada.

Los estudios acerca de este tipo de enfermedades se enfrentan a las siguientes dificultades:

- escasez de pacientes, lo que implica una reducción en la capacidad de generalización y comparación de patrones.
- diferencias en las definiciones fenotípicas, según las regiones geográficas.
- mala documentación e información.
- escasez de financiación, dado su poco interés comercial.

- desconocimiento y/o falta de especialización por parte del personal médico general, derivando en retrasos o ausencias en los diagnósticos y tratamientos.

Por todo lo anterior se antoja imprescindible la colaboración internacional para el abordaje de investigaciones efectivas.

Se estima que hay unas 6.000 enfermedades raras descritas, afectando en Europa a un 7 % de la población [99]. Generalmente suelen estar asociadas a anomalías genéticas estructurales (el 80 % de ellas lo está, según la Organización Europea Para las Enfermedades Raras -EURORDIS- [100]), con lo cual son crónicas, y además en la inmensa mayoría de los casos carecen de tratamientos efectivos.

1.4.2. Detección de variaciones estructurales: genotipo

Las anomalías estructurales a nivel cromosómico hacen referencia a alteraciones de diversos tipos en la estructura del material genético. Dichas anomalías se pueden clasificar en 2 grupos: i) con ganancia o pérdida de material genético (deleciones e inserciones) y ii) sin ganancia ni pérdida de material genético (translocaciones e inversiones).

Tipos de variaciones estructurales (véase la figura 1.9):

- **Deleción:** pérdida de un segmento de material genético en un cromosoma (cuando el área afectada es pequeña se usa el término 'microdeleción'). Responsable de enfermedades tales como el síndrome de Prader-Willi o el síndrome de Angelman, entre otras.
- **Duplicación:** repetición de un fragmento de material genético en un cromosoma, justo a continuación del fragmento original (cuando el área afectada es pequeña se usa el término 'microduplicación'). Responsable de enfermedades tales como la de Charcot-Marie-Tooth tipo I, entre otras.

- **Inversión:** un segmento de material genético en un cromosoma cambia de orientación 180 grados. Normalmente los efectos negativos suelen darse sobre los gametos del individuo.
- **Translocación:** dos segmentos del material genético de dos cromosomas son intercambiados. Existen 2 tipos: i) equilibrada (cuando una región cromosómica cambia de posición en el genoma, pero el número de copias de dicha región se mantiene), cuyos afectados pueden o no presentar síntomas patológicos; y ii) desequilibrada (cuando se produce un cambio en el número de copias), cuyos afectados sí suelen presentar enfermedad.
- **Inserción:** un segmento del material genético de un cromosoma es insertado en otro cromosoma. Si no existe ganancia o pérdida de material cromosómico la persona normalmente será sana.

Cabe destacar que estas anomalías afectan a la estructura del cromosoma en cuanto a la ordenación de sus genes.

Más concretamente, de importancia clínica son las CNVs. Una CNV (del inglés *Copy Number Variation*, o variación en el número de copias) es un segmento de ADN (de al menos 1kb de tamaño) cuyo número de copias en el genoma de un individuo es variable respecto al genoma de referencia. Se trata de anomalías estructurales en forma de deleción o duplicación con variabilidad en la población y con potenciales efectos patológicos [7]. Las CNVs pueden ser, o bien heredadas o bien producidas en el propio individuo (*de novo*).

Existen también CNVs presentes en población sana (en torno al 4,8-9,5 % de la variación del genoma humano [101], como variaciones poblacionales naturales).

Las principales metodologías para detectar CNVs son: *array-Comparative Genomic Hybridization (aCGH)*, *Single Nucleotide Polymorphisms arrays (SNParrays)* y *Next Generation Sequencing* (NGS) [101].

Actualmente, uno de los mayores retos en la investigación es el de asociar estas variaciones estructurales (las CNVs en concreto) con sus efectos: con los fenotipos asociados.

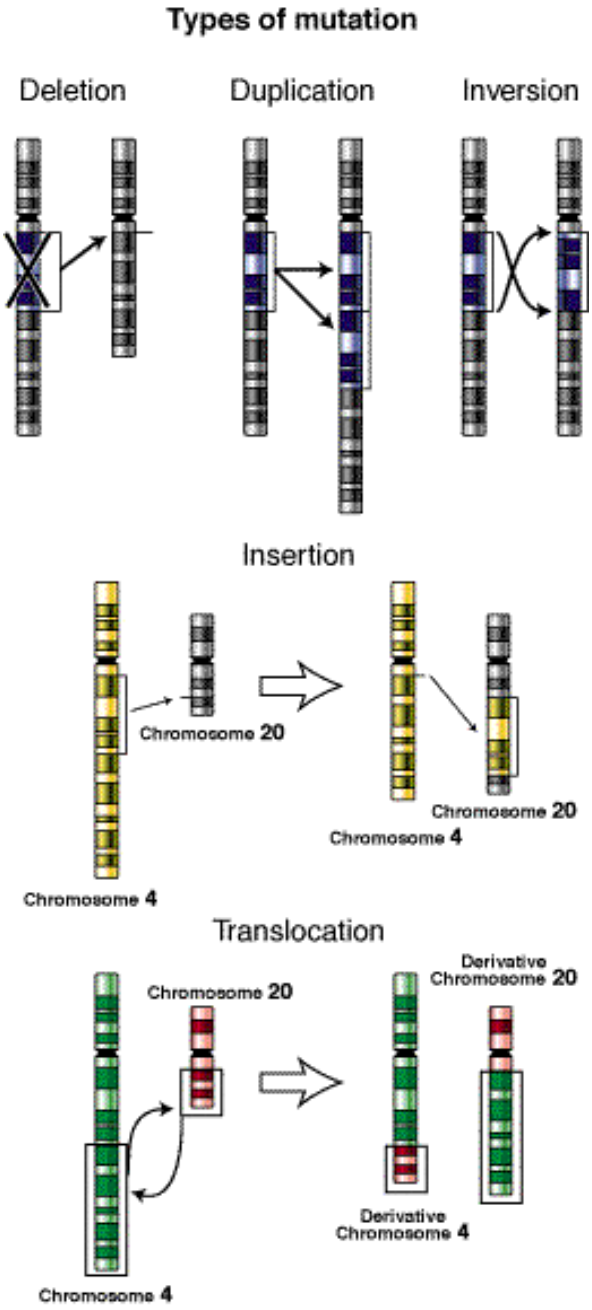


Figura 1.9: **Tipos de variaciones estructurales.** Imagen del *National Institute of Health* de Estados Unidos.

1.4.3. La importancia del fenotipado: la ontología HPO

Definimos un fenotipo como el conjunto de características observables que un individuo presenta como resultado de la interacción entre su genotipo y el medio (morfológicamente, fisiológicamente o en cuanto a comportamiento). En medicina clínica, el fenotipado hace referencia a la clasificación de síntomas y características (normalmente patológicas) de los pacientes [8].

Cada patología humana tiene una combinación específica de características fenotípicas. La manera en la que los clínicos describen las categorías fenotípicas y los términos ha sido, históricamente, caótica y descoordinada; así como las descripciones de los fenotipos humanos incluidas en los artículos de investigación; haciendo muy difícil la comparación entre diferentes estudios. Pero en la era de la medicina personalizada y la Biología de Sistemas, se requiere de un sistema de fenotipado preciso y homogéneo [51]. La estandarización de la anotación fenotípica permite: un almacenamiento automático computerizado, un intercambio transversal y el uso de algoritmos de comparación y análisis sistemáticos.

El fenotipado importa [9], y tiene una importancia que va en aumento a medida que profundizamos en el conocimiento de los desórdenes genéticos. El modelo '1 gen $-$ $>$ 1 patología' está obsoleto y es incierto. Ahora es conocido el hecho de que las mutaciones en el mismo gen pueden llevar a desórdenes diferentes [102] y que síndromes clínicos genéticos considerados como diferentes pueden ser vistos como un desorden de un mismo sistema molecular compartiendo el mismo fenotipo [9]. Así como las mismas enfermedades o fenotipos pueden tener origen poligénico o polisistémico .

Existen algunas comunidades científicas que han desarrollado terminologías para el diagnóstico clínico y molecular: OMIM [103], London Dysmorphology Database (LDDB) [104], POSSUM [105], Orphanet [106], etc. Pero estos vocabularios no están optimizados para el análisis computacional sistémico: no contienen términos apropiadamente anidados, son ambiguas, no con-

tienen los suficientes términos como para permitir un preciso y profundo fenotipado para anomalías genéticas, y algunas de ellas no están disponibles bajo una licencia abierta [10].

La estandarización de los términos fenotípicos permitiría la universalización de su codificación y tanto su procesamiento como su comparación de manera automática. Yendo aún más allá, su estructuración en forma de ontología (una ontología es una definición formal de tipos, propiedades, y relaciones entre entidades que existen para un dominio particular), en la cual un término padre pueda contener términos hijos más específicos en la estructura, así como relaciones con otros términos 'hermanos', daría la oportunidad de clasificar, por capas, y con diferentes grados de precisión, de manera universal los fenotipos y adicionalmente, establecer relaciones de *cercanía* entre ellos [107].

En ese sentido, *The Human Phenotype Ontology* [10, 52, 53] proporciona un vocabulario estandarizado, anidado y controlado para detallar la información fenotípica de una manera no ambigua. Proporcionando una ontología jerárquica, HPO permite la anotación clínica precisa y también el uso de algoritmos computacionales para explotar los datos fenotípicos y sus similitudes semánticas.

HPO es un recurso accesible libremente (http://www.human-phenotype-ontology.org), el cual contiene actualmente un conjunto de más de 13.000 términos, cada uno de los cuales describe una anomalía fenotípica individual codificada con un identificador único (*e.g.* 'HP:0000119' para 'Anomalía del sistema genito-urinario'), y unas 13.326 relaciones entre términos [52]. Los términos están estructurados jerárquicamente mediante relaciones de subclase. Por ejemplo: 'Anomalía de la región periauricular' es una subclase de 'Anomalía del oído externo', que a su vez es una subclase de 'Anomalía del oído' [108].

Robinson *et al.* [10] destaca la importancia de la creación de una base de datos que combine información genética y fenotípica [109] y enfatiza en la idea de que el proyecto HPO proporciona un vocabulario estandarizado para la descripción de fenotipos, lo cual lo hace apropiado para su uso computacional en la investigación clínica en el campo de las anomalías humanas y que puede, además, ser enlazado a información sobre mutaciones en bases da datos específicas sobre *loci*

[110]. Adicionalmente, invita a la creación de grandes bases de datos centralizadas de genotipos-fenotipos para ofrecer una plataforma *web* para el intercambio de datos entre clínicos y permitir el uso y análisis sistemático de datos fenotípicos.

1.4.4. El problema de la relación fenotipo-genotipo

Tras las secuenciaciones masivas de genomas que se han producido desde los años 90, se ha obtenido una importante información acerca de los genes de los diferentes organismos, pero no siempre esto ha ido de la mano de una identificación de sus funciones o implicaciones en el funcionamiento y la apariencia de los seres vivos [111].

La identificación completa de las consecuencias fenotípicas de las CNVs sigue siendo un reto. Es necesario tener en cuenta un amplio número de mecanismos moleculares y genéticos para determinar la relación entre las CNVs y los fenotipos de un individuo:

- ***imprinting*:** El *imprinting* o impronta genética hace referencia al fenómeno por el cual algunos genes son expresados de manera diferente dependiendo del sexo del progenitor.
- **desenmascaramiento de mutaciones recesivas:** en los organismos diploides, las mutaciones pueden o no tener efectos fenotípicos dependiendo del tipo de variación cromosómica y de la combinación que exista entre ambos cromosomas homólogos.
- **disfunción de elementos regulatorios:** actualmente se sabe que existen multitud de elementos regulatorios a nivel genético, molecularmente formados por: ADN, ARN (ARNm, miARN, ...) y proteínas; formando un complejísimo entramado cuya disfunción puede alterar notablemente los efectos de una CNV.
- **herencias complejas:** interacciones, *a priori* impredecibles, entre más de una mutación en el mismo individio.

La integración y comparación a gran escala de fenotipos y genotipos de pacientes con enfermedades genéticas (especialmente de aquellos con enfermedades raras) es esencial para su diagnóstico. Por lo tanto, es de crucial importancia avanzar en la caracterización de las regiones genéticas y los mecanismos moleculares que controlan la expresión fenotípica.

Existen diversas bases de datos disponibles *online* que relacionan variaciones genéticas y fenotipos:

- **PhenIX [112]:** ***Phen****otypic* ***I****nterpretation of* ***eX****omes*, se trata de una herramienta para la obtención de genes candidatos en exomas con asociación a genes humanos de enfermedades mendelianas. Puntúa los genes en base a la patogenicidad predicha en las variaciones así como a la similitud fenotípica de enfermedades asociadas a genes contenidos en dichas variantes, a través de un análisis de *Human Phenotype Ontology* (HPO).

- **Phenomantics [113]:** Puntúa un conjunto de genes dado, relativo al fenotipo de un paciente. Ordena los genes por similitud semántica -computada entre los descriptores fenotípicos asociados con cada gen (procedentes de HPO) y aquellos que describen al paciente-.

- **eXtasy [114]:** Se trata de un procedimiento para la puntuación de SNVs dado un fenotipo. Tiene en cuenta la posible propiedad deletérea de la variante, las predicciones de haploinsuficiencia del gen y la similitud entre el gen dado y otros genes relacionados con ese fenotipo.

- **PHIVE [115] y hiPHIVE [116]:** ***PH****enotypic* ***I****nterpretation of* ***V****ariants in* ***E****xomes* (PHIVE) y su variante **hiPHIVE**, son algoritmos disponibles *online*, que integran cálculos de similitud entre fenotipos a través de enfermedades humanas y ratones modificados genéticamente, evaluando las variantes según la frecuencia del alelo, la patogenicidad y el tipo de herencia. Actúan dentro de la herramienta *Exomiser*.

- **Phevor [117]:** integra información fenotípica, de funcionalidad de los genes y también de las enfermedades; todo ello unido a los datos genómicos personales, para la identificación

de los alelos causantes de las enfermedades. Combina el conocimiento presente en múltiples ontologías biomédicas con los resultados de diversas herramientas de priorización de variantes.

- **Phen-Gen [118]:** se trata de un método que combina los síntomas de las enfermedades que sufren los pacientes con los datos de secuenciación, con la finalidad de identificar los genes causantes de enfermedades raras.

- **OMIM-Explorer [119]:** permite la rápida integración de fenotipos con genotipos para ayudar a los usuarios con diagnósticos diferenciales de enfermedades genéticas, priorización de variantes moleculares y el descubrimiento de nuevas asociaciones gen-fenotipo.

- **Decipher [62]:** se trata de un repositorio abierto a la comunidad científica internacional para compartir y comparar información fenotípica y genotípica de pacientes con desórdenes genómicos de baja prevalencia.

Capítulo 2

Hipótesis y objetivos

No debe haber barreras para la libertad de preguntar. No hay sitio para el dogma en la Ciencia. El científico es libre y debe ser libre para hacer cualquier pregunta, para dudar de cualquier aseveración, para buscar cualquier evidencia, para corregir cualquier error.

J. Robert Oppenheimer

Por lo expuesto anteriormente, esta Tesis Doctoral tiene la hipótesis general de que la modelización de los sistemas biológicos en redes de asociación, como las construidas en este trabajo, ayuda de forma significativa a la caracterización y comprensión de dichos sistemas en diferentes ámbitos: molecular (a nivel de la función de proteínas o genes) y fenotípico (a nivel de patologías o síntomas); asimismo su explotación ayuda a la realización de predicciones que permiten optimizar la experimentación en laboratorio para la identificación de nuevas proteínas funcionales, así como asistir y orientar los diagnósticos clínicos.

Dividiendo la hipótesis principal en cada una de las tres diferentes líneas de investigación de las que se ha compuesto esta Tesis Doctoral, tenemos una subhipótesis más específica para cada uno de los estudios llevados a cabo:

I) El papel funcional de nuevas proteínas aún no caracterizadas, en determinados sistemas moleculares, puede ser predicho a través del análisis de sus relaciones en red con proteínas ya conocidas de dichos sistemas.

II) Existe una relación entre la evolución filogenética de pares de secuencias de proteínas RAS parálogas y su localización en el interactoma humano.

III) La integración en redes de asociación heterogéneas de los datos genéticos (mutaciones) y síntomas (fenotipos) de miles de pacientes con trastornos cromosómicos raros permite identificar de manera sistemática nuevas relaciones significativas genotipo-fenotipo.

En definitiva, esta Tesis Doctoral pretende poner en valor la aplicación de metodologías propias de la Biología Molecular de Sistemas y de la Medicina de Sistemas en la investigación clínica y farmacológica.

Haciendo uso de las estructuras de datos, metodologías y algoritmos detallados en el capítulo 3 se establecen los siguientes objetivos con la finalidad de verificar las hipótesis planteadas:

El objetivo general es la construcción de redes precisas que modelen asociaciones biológicas, así como la implementación de algoritmos que permitan su explotación para el estudio de diversos sistemas a nivel molecular y fenotípico.

Este objetivo se divide en los siguientes subobjetivos:

1) Construir un predictor, mediante el análisis matemático de redes de interacción de proteínas, que permita priorizar estadísticamente aquellas proteínas con más probabilidades de estar involucradas en un sistema molecular dado, con el fin de caracterizarlas funcionalmente.

2) Utilizar las redes de interacción entre proteínas y los árboles filogenéticos de la familia de proteínas parálogas RAS para estudiar la relación entre su evolución molecular y funcional en el interactoma humano.

3) Construir una red *tripartita* a partir de la información más completa posible de pacientes que sufren trastornos cromosómicos raros. Dicha red contendrá las capas: CNVs (mutaciones), pacientes y fenotipos (organizados jerárquicamente en una ontología). Y analizar matemáticamente dicha red con la finalidad de obtener predicciones válidas sobre nuevas relaciones CNV-fenotipo con el objetivo de asistir al diagnóstico clínico en este tipo de pacientes.

Capítulo 3

Materiales y métodos generales

Si supiésemos qué es lo que estamos haciendo... no lo llamaríamos investigación, ¿verdad?

Albert Einstein

En este capítulo se detallan los Materiales y Métodos generales relativos a los estudios que componen la Tesis Doctoral; así como otras técnicas relacionadas.

3.1. Obtención de árboles filogenéticos

En el trabajo desarrollado en esta Tesis Doctoral se ha hecho uso de los árboles filogenéticos calculados para la familia de proteínas RAS en 24 organismos. Esta información fue obtenida de Diez *et al.* [29] en una colaboración entre la Universidad de Kioto y la Universidad de Málaga. Se extrajeron los datos relativos a esta familia de proteínas en *Homo sapiens* (35 parálogos), se transformaron las relaciones filogenéticas (en formato *Newick Standard*, véase el capítulo 1.2) en distancias entre pares y se compararon con las distancias en las redes de interacción de proteínas. Este tipo de análisis combinados permite establecer relaciones entre los cambios moleculares y funcionales de las proteínas e incluso focalizar los estudios en pequeños cambios -o similitudes- moleculares que puedan ser responsables de cambios -o similitudes- funcionales. El estudio completo, aplicado a la familia de proteínas parálogas RAS en *Homo sapiens*, se muestra en el capítulo 5.

3.2. Análisis de redes de interacción de proteínas

Muchas son las bases de datos y métodos de predicción de interacciones entre proteínas disponibles. Según su naturaleza podemos clasificar las principales de ellas como sigue:

- Interacciones identificadas experimentalmente: BioGRID [38], Database of Interacting Proteins (DIP) [42], HPRD [39], IntAct [40], KEGG [19], MIPS/MPact [43, 44], Molecular Interaction Database (MINT) [41] y Reactome [24].
- Predicciones bioinformáticas: CODA [25], GECO [26], HIPPIE [27], iHop [120] y Predictome [121].
- Combinación de varias fuentes (se detallan a continuación): PINA [22], STRING [23] e iRef [21].

  - **PINA** [22] (Protein Interaction Network Analysis platform) contiene información integrada procedente de 6 bases de datos públicas curadas manualmente: BioGRID [38], DIP [42], HPRD [39], IntAct [40], MINT [41] y MIPS/MPact [43, 44].

  - **STRING** [23] está formada por la unión ponderada de varias fuentes de datos: BIND [37], Biocarta, BioCyc, BioGRID [38], DIP [42], GO [20], HPRD [39], IntAct [40], KEGG [19], MINT [41] y Reactome [24]. Además de proporcionar una puntuación global para cada interacción (según su grado de fiabilidad), también diferencia la naturaleza de todas estas interacciones, dividiendo las puntuaciones en canales según su procedencia (tipo de evidencia). A continuación se muestra el listado de canales de los que hace uso STRING:

  nscore - por vecindad en el genoma (a partir del conteo de nucleótidos entre genes).

  fscore - por fusión (genes que se fusionan en un solo marco abierto de lectura en otras especies).

  pscore - por coaparición de perfiles (similitud en los patrones de presencia/ausencia de ge-

nes).

hscore - por homología (grado de homología con interactores conocidos).

ascore - por coexpresión (patrones de similaridad en expresión de mRNA).

escore - experimental (datos experimentales en laboratorio, *e.g.* coinmunoprecipitación, Y2H, etc.).

dscore - puntuación heredada de otras bases de datos.

tscore - por minería de textos (cocitación de nombres de genes/proteínas en publicaciones científicas).

- **iRef** [21] contiene información integrada procedente de 10 bases de datos públicas: BIND [37], BioGRID [38], CORUM [122], DIP [42], HPRD [39], IntAct [40], MINT [41], MIPS/MPact [43, 44] y OPHID [123].

Estas bases de datos fueron extensamente analizadas y evaluadas en el trabajo desarrollado en esta Tesis Doctoral, con el fin de modelar redes de interacción de proteínas consistentes. En cada caso concreto se tuvieron en cuenta diferentes factores para la selección de las fuentes de datos, tales como: la fiabilidad, el origen y la cantidad de la información o la relevancia dentro del contexto biológico de estudio.

3.2.1. Algoritmos de medida de distancias en redes

Tal como se mencionó en la Introducción general (capítulo 1.3.1), se requiere de métodos de cálculo de *distancias* en red, con la finalidad de obtener información acerca de la similitud de contextos funcionales entre pares de proteínas en el interactoma, fundamental para los estudios aquí presentados. A continuación se detallan, minuciosamente, las principales opciones consideradas en este trabajo para medir dichas *distancias*, así como la selección final de las medidas usadas. Estas medidas se presentan divididas en 3 grupos: **a)** algoritmos de búsqueda de los caminos más cortos, **b)** algoritmos probabilísticos de difusión y **c)** comparaciones de perfiles.

a) Algoritmos de búsqueda de los caminos más cortos

La opción más intuitiva es la de la búsqueda del camino más corto entre cada par de nodos. Este tipo de análisis, si se aplica a toda la red, da como resultado una matriz (nxn) de distancias entre cada par de nodos del grafo (n nodos), de forma que se pueden ordenar las distancias de mayor a menor para posteriores análisis o también comparar de manera directa dichas distancias con otras métricas que impliquen a esos mismos pares de entidades (*e.g.*, si se están analizando proteínas, se podrían comparar con las distancias en el árbol filogenético).

El científico de la computación holandés Edsger Wybe **Dijkstra** [124] propuso su solución en 1959 mediante el algoritmo homónimo. Computacionalmente se trata de un problema de complejidad *P*. Dentro de la *Teoría de la Complejidad Computacional*, la clase de problemas *P* engloba a aquellos algoritmos de decisión que pueden ser resueltos en una máquina determinista secuencial en un período de tiempo polinomial en proporción a la cantidad de datos de entrada. Téngase en cuenta que los tiempos de ejecución y el conocimiento de la complejidad algorítmica son claves para trabajar con datos de entrada tan abundantes como son los procedentes de redes biológicas.

Mediante el algoritmo de Dijkstra podemos obtener las rutas más cortas entre los nodos de nuestro grafo de la siguiente manera:

Teniendo un grafo de n nodos, siendo x el nodo inicial, un vector D de tamaño n guardará al final del algoritmo las distancias desde x al resto de los nodos mediante las siguientes reglas:

1. Se inicializan todas las distancias en D con un valor infinito, ya que son desconocidas al principio, exceptuando la de x que se debe colocar a 0 debido a que la distancia de x a x será 0.
2. Sea a = x (tomamos a como nodo actual).
3. Recorremos todos los nodos adyacentes de a, excepto los nodos marcados, llamaremos a estos vi.

4. Si la distancia desde x hasta vi guardada en D es mayor que la distancia desde x hasta a, sumada a la distancia desde a hasta vi; esta se sustituye con la segunda nombrada, esto es: $si(Di > Da + d(a, vi))\ entonces\ Di = Da + d(a, vi)$

5. Marcamos como completo el nodo a.

6. Tomamos como próximo nodo actual el de menor valor en D y volvemos al paso 3 mientras existan nodos no marcados.

Una vez terminado el procesamiento del algoritmo, D estará completamente lleno.

El problema de este algoritmo es que hay que repetirlo para cada nodo de la red, tratándose de un examen exhaustivo del grafo, de alta demanda computacional, lo cual lo convierte en un inconveniente cuando trabajamos con redes grandes que implican cientos de miles de nodos.

Dando un paso más allá tenemos los algoritmos que calculan la distancia más corta haciendo uso de heurísticas, es decir, sin recorrer todo el espectro de posibilidades y ahorrando así una ingente cantidad de tiempo de computación, haciendo más viable el análisis. Se basan en estimaciones, eliminando las opciones que *a priori* parecen no llevar al resultado óptimo y reduciendo así el espacio de posibilidades a analizar. No se puede garantizar al cien por cien que la solución sea la mejor, pero con un alto grado de probabilidad será una de las mejores, si no efectivamente la mejor de ellas; es decir, uno de los caminos más cortos. Entre estos algoritmos destaca el algoritmo $\boldsymbol{A}^*$ [125], propuesto en 1968, y con diferentes variantes: heurísticas, exhaustivas y mixtas. Este algoritmo sigue requiriendo de una gran cantidad de tiempo de ejecución, así como de alta capacidad de memoria durante el análisis, ya que va almacenando dinámicamente los siguientes posibles nodos en cada estado del proceso.

El funcionamiento del algoritmo A^* es el siguiente:

```
ABIERTOS := [INICIAL] //inicialización
CERRADOS := []
f'(INICIAL) := h'(INICIAL)
repetir
        si ABIERTOS = [] entonces FALLO
        si no // quedan nodos
                extraer MEJORNODO de ABIERTOS con f' mínima
                  // cola de prioridad
                mover MEJORNODO de ABIERTOS a CERRADOS
                si MEJORNODO contiene objetivo entonces
                  SOLUCIONE-NCONTRADA := TRUE
                si no
                  generar SUCESORES de MEJORNODO
                  para cada SUCESOR hacer TRATAR-SUCESOR
hasta SOLUCION-ENCONTRADA o FALLO
```

Donde *ABIERTOS* será el vector de nodos por resolver y *CERRADOS* el de nodos ya resueltos. La función de evaluación es $f(n) = g(n) + h'(n)$, donde $h'(n)$ representa el valor heurístico del nodo a evaluar desde el actual, n, hasta el final, y $g(n)$ el coste real del camino recorrido para llegar a dicho nodo n.

Básicamente, resumiríamos el algoritmo así: Un nodo es extraído de la estructura abierta, el cual se somete a comprobaciones de si es nodo objetivo, factor que determina el comportamiento del algoritmo: si la evaluación resulta ser afirmativa, se obtiene el resultado, en caso contrario, se expande este nodo analizando sus hijos mediante la función de evaluación y procesándolos y ordenándolos nuevamente en la estructura temporal.

Otro método adicional es el algoritmo ***Floyd-Warshall*** [126], descrito en 1959 por Bernard Roy, que se basa en programación dinámica, una técnica establecida en 1953, utilizada para optimizar problemas complejos mediante la discretización y secuenciación de los mismos, es decir, dividiéndolos en subproblemas superpuestos y en subestructuras de datos. Se emplean soluciones óptimas a subproblemas para encontrar la solución óptima global. *Floyd-Warshall* encuentra el camino más corto entre todos los pares de nodos de un grafo en una única ejecución, y lo hace de forma paulatina mediante un bucle. Su funcionamiento es el siguiente:

Definimos el $caminoMinimo(i, j, k)$ (CM) de forma recursiva:

$$CM(i, j, k) = min(CM(i, j, k - 1), CM(i, k, k - 1) + CM(k, j, k - 1));$$

$$CM(i, j, 0) = 1;$$

Siendo i y j los nodos entre los cuales se quiere calcular la distancia y k el número de aristas seleccionadas. En cada punto del análisis se selecciona la opción mínima entre las dos: i) o bien existe un camino mínimo que hace uso de las aristas desde 1 hasta k ii) o bien existe un camino que va desde i hasta k y desde k hasta j que es más corto.

Los métodos descritos hasta el momento tienen algunas desventajas: las distancias que devuelven son discretas (números naturales), de manera que no podemos realizar un ajuste fino a la hora de interpretar los resultados ni se refleja con exactitud la topología de la red. Se trata de aproximaciones toscas. Por ejemplo, el nodo 3 y el nodo 8 en la figura 3.1 tienen la misma distancia directa que el nodo 8 y el nodo 11. Sin embargo, el significado de ambas relaciones es muy diferente debido a la distribución de agrupaciones de elementos en el contexto de la topología de la red. Para evitar estos problemas se puede recurrir a algoritmos basados en probabilidades y difusión de la señal, los cuales sí tienen en cuenta los contextos topológicos derivados de la estructura de la red y devuelven valores más precisos (números decimales).

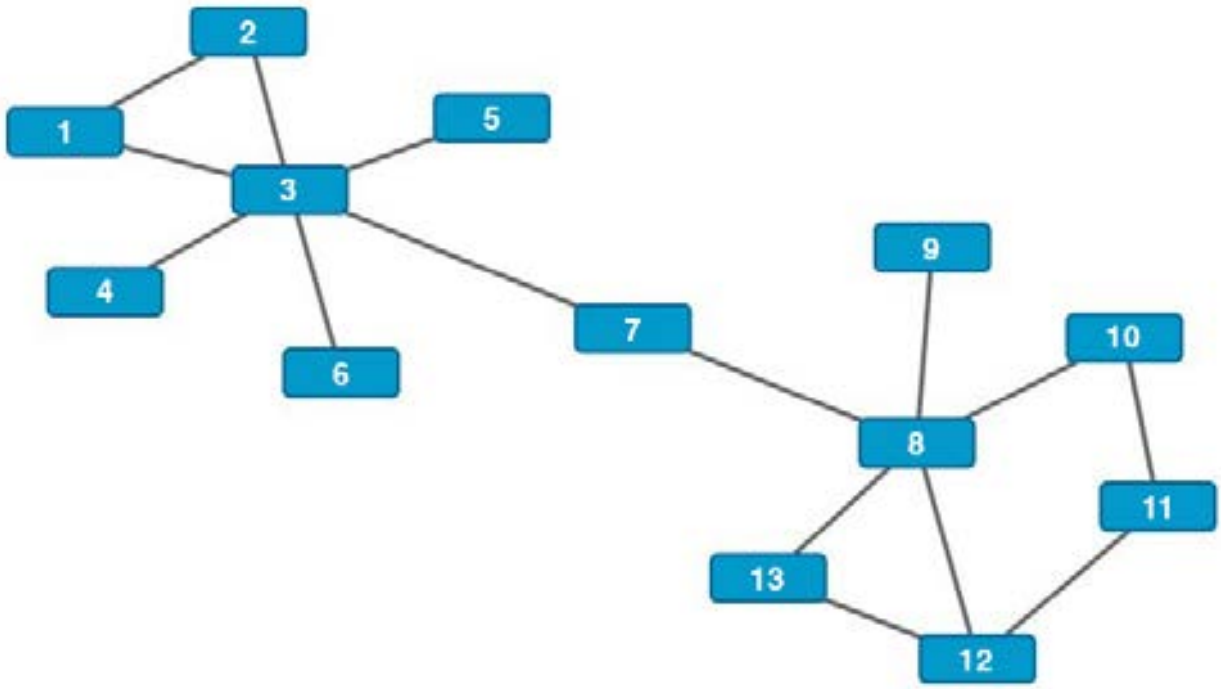


Figura 3.1: **Ejemplo de grafo.** Los rectángulos representan los nodos del grafo y las líneas son las aristas (interacciones entre ellos).

b) Algoritmos probabilísticos de difusión

Los algoritmos de análisis de redes con base en probabilidades de difusión son más sofisticados. No sólo miden las distancias, sino que también tienen en cuenta la topología de la red y toman en consideración las rutas indirectas entre pares de nodos. Así mismo, tienen la propiedad de incrementar o decrementar sus valores de distancia dependiendo del número de caminos que conecten 2 nodos (de la redundancia en la conectividad), cosa que mediante los cálculos del camino más corto no ocurre y conservan siempre un mismo valor, aunque se añadan o retiren rutas alternativas.

Estos métodos se basan en la suma de los pesos de los caminos entre cada par de nodos computándolos en un espacio de características *n-dimensional*, analizando las proyecciones que se generan. Matemáticamente, se mapean los nodos (*e.g.* proteínas) en ese espacio de características multidimensional y se computan las diferencias entre los vectores generados (proyecciones). El resultado final es una matriz que contiene la evaluación de dichas comparativas y que puede ser considerada como una matriz de similitud entre nodos. Los algoritmos de este tipo tienen en cuenta la bidireccionalidad de la red, haciendo uso de medias de difusión, y la mayoría de ellos devuelven los mismos valores de distancia para los pares, independiendemente de cual sea el sentido de la conexión: a-b = b-a.

Otro punto muy importante a favor de estos métodos es que tienen en cuenta el efecto que producen en la topología de la red los nodos promiscuos (*hubs*), y las múltiples conexiones (*switches*), amortiguando los posibles artefactos que puedan introducir en los resultados de los análisis.

Podemos destacar los siguientes algoritmos de este tipo[1]:

1) *Random Walk* (RW): Potente algoritmo de medida de distancias en red introducido en 1905 por el matemático inglés Karl Pearson, que se basa en el análisis estadístico de la secuencia aleatoria de pasos que se tomarían partiendo de un nodo origen hasta la llegada a un nodo destino, pasando en cada iteración a uno de los nodos adyacentes. El algoritmo se ejecuta reiteradamente y el 'tiempo' que se pasa en un determinado nodo con respecto al numero total de iteraciones determina el valor probabilístico de afinidad de ese nodo con respecto al nodo origen. Existe una modificación de este algoritmo que reinicia el estado, devolviéndolo al nodo inicial, cada x iteraciones, de forma que no se aleje excesivamente desvirtuando el resultado. Esta variante se conoce como Random Walk With Restart (RWR) [127–130].

Formalmente definiríamos el RWR con la siguiente ecuación:

$$p(t+1) = (1-r)Wp(t) + rp(0)$$

Siendo W la matriz de adyacencia normalizada del grafo, pt el vector donde el elemento i contiene la probabilidad de estar en el nodo i en el paso t, r la probabilidad de reinicio y $p0$ la probabilidad inicial (igualdad de probabilidades para todos los nodos, es decir 1 dividido entre el numero total de nodos).

[1]Definición de las matrices utilizadas para representar las redes y sus posteriores derivadas durante los análisis:

Matriz de adyacencia (A): contiene un número de filas y columnas idéntico al número de nodos del grafo, presentando un *1* en aquellas posiciones en las cuales exista una arista uniendo los nodos de esa posición y un *0* en caso contrario.

Matriz identidad (I): contiene un número de filas y columnas idéntico al número de nodos del grafo y cumple la propiedad de ser neutra en un producto de matrices. Está formada por valores *0* en todas las posiciones excepto en la diagonal, donde presenta valores *1*.

Matriz de grado (D): contiene un número de filas y columnas idéntico al número de nodos del grafo y forma una matriz diagonal, presentando valores *0* en las celdas no diagonales y los valores de *grado* de cada vértice en la diagonal.

Matriz laplaciana (L): Se obtiene a partir de la matriz de grado y la matriz de adyacencia de la siguiente manera: $L = D - A$.

Matemáticamente hablando, el *Random Walk* es una cadena finita de Markov que es *reversible en el tiempo* [131, 132]. De hecho, no existe mucha diferencia entre la teoría de RW aplicada a grafos y los modelos finitos de cadenas de Markov; a efectos prácticos las cadenas de Markov reversibles en el tiempo pueden ser vistas como procesos de *Random Walk* aplicados a grafos no dirigidos.

Una desventaja de este método es el hecho de que devuelve una matriz no simétrica: la distancia calculada para el par de nodos *a*-*b* no es la misma que para el par *b*-*a*, y esto complica la interpretación de los resultados.

2) *Random forest kernel* (RF): Este algoritmo del subtipo *kernel* surge de la enumeración de las raíces formadas por los caminos en el grafo y mide la 'accesibilidad relativa' entre nodos [133]. También tiene una interpretación en términos de probabilidades de alcanzar un nodo en un paseo aleatorio (*random walk*) con un número aleatorio de pasos [134]. Los resultados son simétricos. La fórmula del *Random forest kernel* es:

$$RF = (I + L)^{-1}$$

Siendo I la matriz de identidad y L la matriz Laplaciana del grafo.

3) *von Neumann diffusion Kernel* (VN): Este *kernel* enumera todos los caminos entre 2 nodos y penaliza los más largos. Sus resultados son simétricos [135]. Su fórmula es la siguiente:

$$VN = \sum_k \alpha^k A^k = (I - \alpha A)^{-1}$$

α^k es el factor de penalización (siendo k la longitud del camino) y el *kernel* se define para $0 < \alpha < \rho^{-1}$, siendo ρ el radio espectral de A.

4) *Laplacian Exponential Diffusion Kernel* (DK) [136, 137]:

El funcionamiento del método DK es el siguiente: Siendo A la matriz de adyacencia del grafo, y D la matriz de grado en el espacio de características, de forma que $D[i][i]$ es el grado del nodo i; entonces se construye la matriz laplaciana de la forma: $L = D - A$. Una vez obtenido esto, el resultado del *Laplacian Exponential Diffusion Kernel* será:

$$DK = exp(-\beta L)$$

Donde β (siempre mayor que 0) controla la magnitud de la difusión y es el parámetro a ajustar para dotar de mayor precisión a los resultados. Con un valor de β lo suficientemente pequeño, el método puede ser considerado como un *Random Walk* 'vago', con probabilidades β de pasar de un nodo a cualquiera de los vecinos, y además con la probabilidad de permanecer en el mismo nodo de $1 - d(i)\beta$ (siendo $d(i)$ el grado del nodo i) [138].

En los trabajos presentados en esta Tesis Doctoral se hicieron diversos tests de validación con la finalidad de ajustar el parámetro β de manera óptima, y finalmente se fijó en un valor de 0,02 por haber sido el que mejores resultados proporcionaba.

Tras la ejecución del algoritmo se obtiene una matriz *n x n* simétrica (siendo *n* el número de nodos en la red) conteniendo las distancias entre cada par de nodos.

5) *Commute Time Diffusion Kernel* (CT):

Este método, también del subtipo *kernel*, se basa en la computación de la media del número de pasos que un caminante aleatorio necesitaría para ir de un nodo a otro y volver [129]. También tiene una interpretación en términos de redes eléctricas, siendo comparable a la resistencia efectiva entre 2 nodos [46]. Y como un tercer, y quizás más claro ejemplo, sería como considerar la red como un entramado de tuberías, abrir un grifo en el nodo origen y obtener como resultado la cantidad de agua que llega al nodo destino. Se define matemáticamente como:

$$CT = L^{+}$$

Es decir, la pseudoinversa de la matriz laplaciana (L). Los resultados son simétricos.

Tal como se ha indicado, los métodos basados en probabilidades de difusión tienen en consideración los efectos de la topología de la red a la hora de medir las distancias entre los nodos, como se puede apreciar en la figura 3.2. La medida mediante los métodos de distancia más corta devolvería el mismo valor en los 3 casos mostrados en la figura (distancia más corta = 2 aristas), sin embargo el algoritmo probabilístico de difusión (*CT* en este caso) es capaz de distinguir entre las 3 situaciones planteadas, y proporciona un valor más alto para la conexión entre *x* e *y* en el escenario C que en el A o el B.

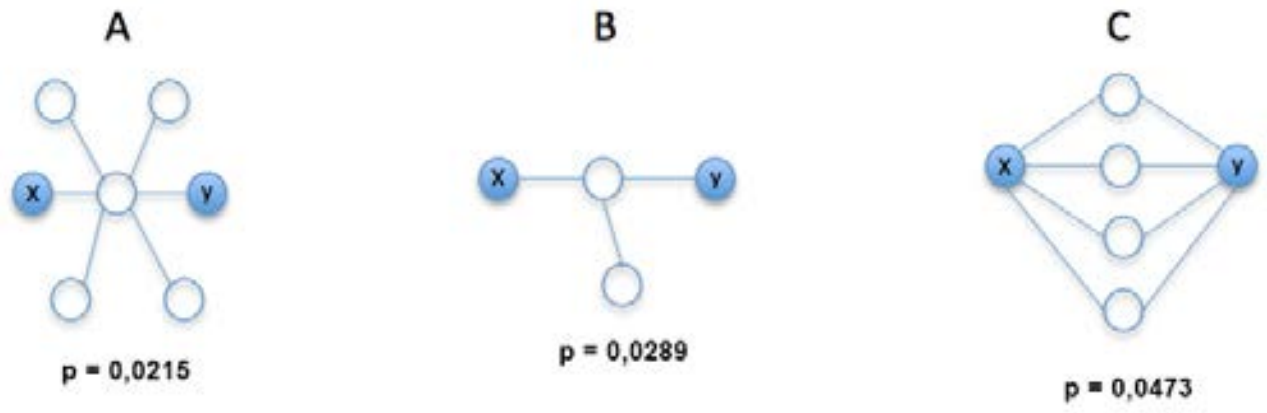


Figura 3.2: **Ejemplo de diferentes topologías de red conectando 2 nodos (*x* e *y*).** El valor de distancia (p), calculado mediante el algoritmo CT, se muestra bajo cada modelo de red. A) Un nodo altamente conectado (*hub* o nodo promiscuo) conecta *x* e *y*. B) El nodo *x* y el nodo *y* se encuentran conectados a través de un nodo intermedio con baja conectividad. C) El nodo *x* y el nodo *y* están conectados a través de muchos caminos. *Imagen adaptada de Köhler et al. [139].*

Cabe destacar que podemos dividir los 5 algoritmos enumerados en esta sección entre los no paramétricos: *Commute Time Diffusion Kernel* y *Random Forest Kernel*; que presentan resultados robustos aunque menos ajustados y los paramétricos: *Random Walk with Restart*, *von Neumann diffusion Kernel* y *Laplacian Exponential Diffusion Kernel*; que pueden ser más precisos si son

bien ajustados o mucho menos precisos si no se dispone de un conjunto de datos de prueba para establecer previamente sus parámetros óptimos.

Los métodos *Commute Time Diffusion Kernel* y *Laplacian Exponential Diffusion Kernel* (implementado con parámetros optimizados), mostraron los mejores rendimientos en un estudio comparativo reciente [46] y por ello, han sido los que se han seleccionado para su uso en esta Tesis Doctoral.

Como último dato, se debe tener en cuenta que los métodos basados en distancias directas representan los valores de cercanía en la red de una manera intuitiva según el número de pasos que haya que dar para llegar de un nodo a otro, y por lo tanto, un valor menor implica una mayor cercanía. Sin embargo, en los métodos basados en probabilidades ocurre a la inversa y las distancias se distribuyen entre los valores 0 y 1, siendo 0 la mayor lejanía y 1 la mayor cercanía en el contexto de interacciones.

c) Algoritmos de comparación de perfiles

La última medida de similitud en red a considerar es la comparación de perfiles entre cada par de nodos, es decir, la valoración de las diferencias que existen entre el vector formado por todas las distancias desde el nodo A al resto de nodos de la red y el vector formado por todas las distancias desde el nodo B al resto de nodos de la red; esto da una orientación sobre lo parecidos que son los contextos de interacción en la red de ambos nodos y por lo tanto una medida más de similitud.

Como ejemplo analizaremos el método de *Distancia Euclídea* entre perfiles: Una distancia euclídea no es más que la distancia 'ordinaria' que podemos medir entre 2 puntos en un espacio euclídeo, deducido a partir del Teorema de Pitágoras. En nuestro caso, extendemos este concepto para aplicarlo a las distancias en red, de forma que dada una matriz de distancias (calculada mediante cualquiera de los métodos citados hasta el momento) podemos obtener la distancia euclídea entre 2 proteínas (entre sus 2 perfiles) simplemente seleccionando los 2 vectores correspondientes

a estos nodos en la matriz (las filas que los representan) y aplicando la fórmula general de distancia euclídea:

$$d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 ..... (z_2 - z_1)^2}$$

Siendo P_1 y P_2 los 2 perfiles a comparar compuestos respectivamente por los elementos: $x1, y1, ...z1$ y $x2, y2, ...z2$.

Como puede apreciarse, este método parte de los anteriores por lo que es considerado en la literatura como una *meta-métrica*.

Algunas de las técnicas de análisis filogenético y de análisis de distancias en redes de interacción de proteínas descritas en esta sección fueron aplicadas, en los trabajos llevados a cabo durante esta Tesis Doctoral, a la familia de proteínas RAS con la finalidad de descubrir la relación entre su evolución molecular y funcional (véase el capítulo 5). Los resultados fueron publicados y forman parte de la colección de artículos que avalan esta Tesis Doctoral (véase el capítulo 5.3.6).

3.2.2. Algoritmos de medidas de similitud en redes heterogéneas

Existe un determinado grupo de redes en las cuales, debido a su naturaleza, no es posible aplicar medidas de similitud como las descritas hasta el momento. Se trata de redes heterogéneas, en las cuales los nodos no son todos de la misma naturaleza y las distancias o perfiles no pueden ser calculados de igual manera, puesto que las relaciones no se establecen de la misma forma, al contrario de lo que ocurría en las redes homogéneas.

Un subtipo de redes heterogéneas es el de aquellas formadas por capas. En ese caso podemos tener: redes *bipartitas*, redes *tripartitas*, etc.; donde los nodos de una capa se conectan exclusivamente con los de otra capa (véase la figura 3.3).

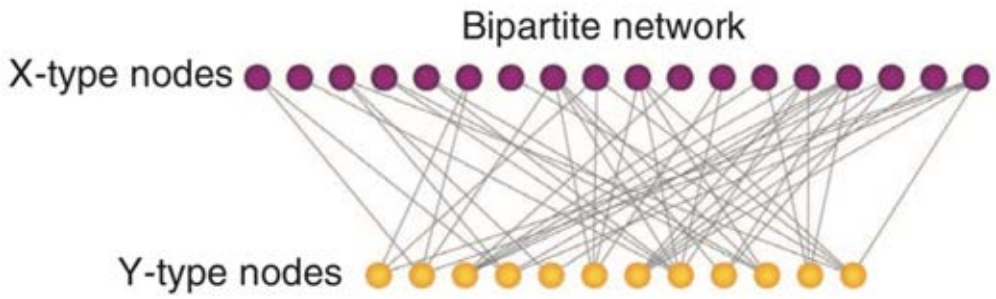


Figura 3.3: **Ejemplo de red *bipartita*.** Gráfico extraído de Bass *et al.* [67].

En estos casos existe un conjunto de algoritmos, basados en probabilidades, destinados al análisis de conexiones entre nodos de diferentes capas en redes heterogéneas. En la figura 3.4 se muestran algunos de ellos.

Algunos de los algoritmos descritos en este capítulo fueron aplicados durante un estudio presentado más adelante en esta Tesis Doctoral (véase el capítulo 6) para el análisis de la base de datos DECIPHER [62], la cual contiene información a 3 niveles: pacientes, variaciones genómicas y fenotipos; permitiendo crear una red heterogénea *tripartita*. Los resultados fueron publicados y forman parte de la colección de artículos que avalan esta Tesis Doctoral (véase el capítulo 6.3.4).

The **Jaccard** index is the proportion of shared nodes between A and B relative to the total number of nodes connected to A or B.

$$J_{AB} = \frac{|N(A) \cap N(B)|}{|N(A) \cup N(B)|}$$

The **Simpson** index is the proportion of shared nodes relative to the degree of the least-connected node.

$$S_{AB} = \frac{|N(A) \cap N(B)|}{\min(|N(A)|, |N(B)|)}$$

The **geometric** index corresponds to the product of the proportion of shared nodes between A and B.

$$G_{AB} = \frac{|N(A) \cap N(B)|^2}{|N(A)| \cdot |N(B)|}$$

The **cosine** index is the geometric mean of the proportions of shared nodes between A and B.

$$C_{AB} = \frac{|N(A) \cap N(B)|}{\sqrt{|N(A)| \cdot |N(B)|}}$$

The **Pearson correlation coefficient** is the correlation between the interaction profiles of A and B.

$$PCC_{AB} = \frac{|N(A) \cap N(B)| \cdot n_y - |N(A)| \cdot |N(B)|}{\sqrt{|N(A)| \cdot |N(B)| \cdot (n_y - |N(A)|) \cdot (n_y - |N(B)|)}}$$

The **hypergeometric** index is the log-transformed probability of having an equal or greater interaction overlap than the one observed between A and B.

$$H_{AB} = -\log \sum_{i=|N(A) \cap N(B)|}^{\min(|N(A)|, N(B)|)} \frac{\binom{|N(A)|}{i} \cdot \binom{n_y - |N(A)|}{|N(B)| - i}}{\binom{n_y}{|N(B)|}}$$

The connection specificity index (**CSI**) is defined as the fraction of X-type nodes that have an interaction profile similarity with A and B that is lower than the interaction profile similarity between A and B itself.

$$CSI_{AB} = 1 - \frac{\#\text{nodes connected to A or B with PCC} \geq PCC_{AB} - 0.05}{n_y}$$
$$= \frac{\#\text{nodes connected to A and B with PCC} < PCC_{AB} - 0.05}{n_y}$$

Figura 3.4: **Algoritmos de análisis de similitud en redes heterogéneas.** Siendo A y B nodos de diferentes capas. $N(A)$ el grado del nodo A: el número de nodos con el que interactúa. Y n_y el número total de nodos en la capa intermedia. Imagen extraída de Bass *et al.* [67].

3.3. Creación de predictores de nuevas proteínas implicadas en procesos moleculares en base a análisis de red y validación mediante el método LOO y curvas ROC

En uno de los trabajos presentados en esta Tesis Doctoral (capítulo 4) se ha desarrollado la metodología necesaria para la construcción de un predictor computacional cuya finalidad es la de descubrir nuevas proteínas implicadas en determinados procesos moleculares.

Dicho predictor ha sido desarrollado haciendo uso de técnicas de análisis *kernel* (véase el capítulo 3.2.1) sobre redes de interacción de proteínas (véase el capítulo 1.3). Las *distancias* obtenidas por los análisis para cada proteína del interactoma humano en relación a un *set de referencia* en la red, formado por un conjunto de proteínas que se saben implicadas en el proceso molecular de estudio, proporcionan una medida de la semejanza de los contextos funcionales entre esa proteína y el *set de referencia*, que tras los pertinentes análisis estadísticos de significancia, puede ser utilizada como valor de la predicción de pertenencia a ese grupo funcional.

Por lo tanto, con la aplicación de la metodología descrita ya se dispondría de los datos de predicciones para todo el interactoma: un valor para cada proteína, en relación a la estimación de su pertenencia al sistema molecular de estudio. Esta metodología ha sido posteriormente validada mediante métodos *leave one out* (LOO) y curvas ROC (*Receiver Operating Characteristic*). Los principios de esta validación se detallan a continuación.

Los contextos biológicos en los cuales dichos predictores fueron aplicados y sus resultados se muestran en el capítulo 4. Parte de este trabajo contribuyó a una publicación relacionada con el proceso molecular de la angiogénesis [14] que se adjunta en el capítulo 4.4.3.

3.3.1. Validación cruzada *Leave One Out* (LOO)

El método LOO (*Leave One Out*, o *dejar uno fuera*) es un tipo concreto de validación cruzada. La validación cruzada (*cross-validation* en inglés) es una técnica usada para la evaluación de resultados estadísticos que consiste en repetir un experimento (o predicción) con un conjunto de datos conocido (sobre el que se sabe cual debería ser el resultado) dividiéndolo en 2 subconjuntos: *datos de referencia* (que servirá para construir el modelo -predictor-) y *datos problema* (con los que ejecutaremos el predictor como si fuesen datos desconocidos), de manera que el sistema tomará los primeros como datos de partida del problema y tratará de clasificar a los segundos. El proceso es repetido varias veces, dejando cada vez fuera del conjunto *datos de referencia* a un subconjunto diferente del mismo (y marcándolo como *datos problema*) hasta completar el total de datos iniciales. Finalmente, se calcula la media de las medidas de evaluación sobre las diferentes particiones. De esta manera, y puesto que ya inicialmente todos los datos utilizados tenían resultados conocidos, se puede estimar la precisión del modelo predictivo [140]. En el caso de LOO, la validación se lleva a cabo etiquetando como *datos problema* únicamente a uno de los elementos del grupo *datos de referencia* en cada ejecución (véase la figura 3.5).

En los estudios llevados a cabo en este trabajo, los 'datos de referencia' hacen alusión a las proteínas que se sabe forman parte del sistema molecular de estudio, y los 'datos problema' al resto de proteínas del interactoma (o en su defecto a un subconjunto seleccionado del interactoma) y las puntuaciones dadas a estas últimas serán los valores del predictor. Para la validación mediante LOO, tal como se ha explicado, únicamente se utilizan las proteínas pertenecientes a los 'datos de referencia', por tener valores conocidos; permutando su pertenencia a este grupo, tal como se ha explicado.

Finalmente, los datos de la validación LOO pueden ser representados mediante una curva ROC (véase la siguiente sección).

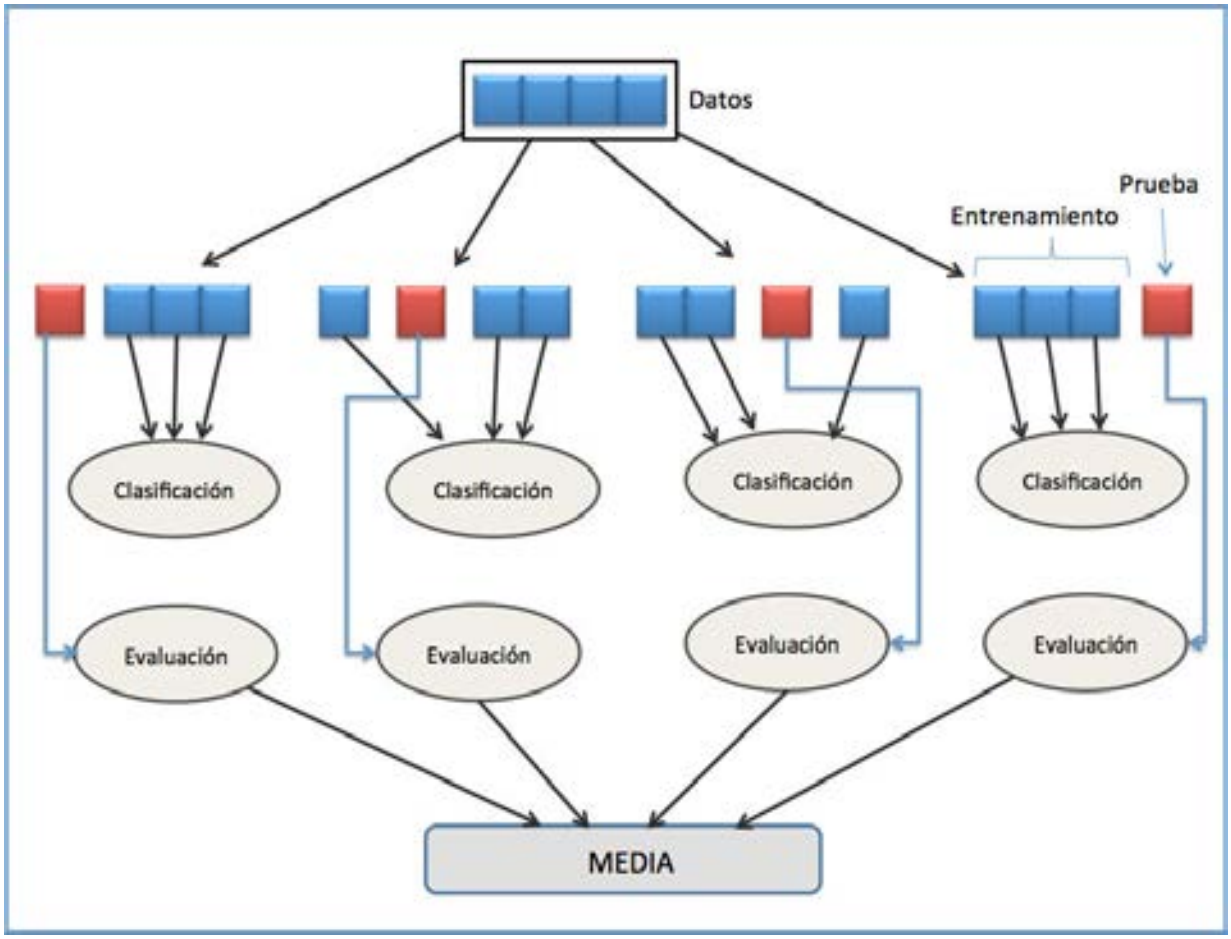


Figura 3.5: **Ejemplo de aplicación de validación cruzada LOO.** Del conjunto de datos conocido y usado como referencia, se extrae en cada iteración uno de los elementos y se evalúa (como si fuese desconocido) mediante el predictor, usando el resto como referencia. Tras repetir este proceso con cada uno de los elementos, se puede obtener una media de la calidad del predictor.

3.3.2. La curva ROC y el AUC

La denominada *curva ROC* (*Receiver Operating Characteristic*, o Característica Operativa del Receptor) es una representación gráfica de un conjunto de datos, dentro del marco de la Teoría de detección de señales, que permite estimar la calidad de una predicción. Para ello se representan la **sensibilidad** frente a la **especificidad** de los resultados del predictor, según la variación del umbral de discriminación. O también, la razón de verdaderos positivos frente a la razón de falsos positivos (véase la figura 3.6).

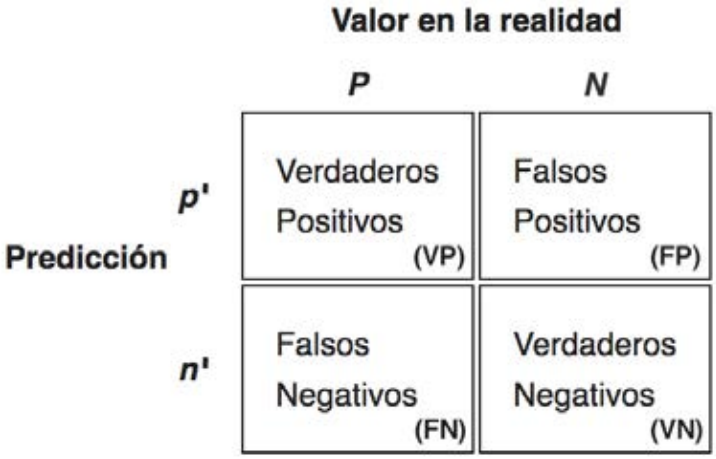


Figura 3.6: **Cuadro de posibles valores devueltos por un predictor.**

Tenemos los siguientes parámetros:

- Siendo:

  P = positivos reales; dentro del conjunto de datos de entrenamiento.

  N = negativos reales; dentro del conjunto de datos de entrenamiento.

  p' = positivos predichos; elementos que el sistema considera como válidos en su predicción.

n' = negativos predichos; elementos que el sistema considera como no válidos en su predicción.

VP = verdaderos positivos; elementos que el sistema considera como válidos en su predicción y que realmente lo son.

VN = verdaderos negativos; elementos que el sistema considera como no válidos en su predicción y que realmente no lo son.

FP = falsos positivos; elementos que el sistema considera como válidos en su predicción y que realmente no lo son.

FN = falsos negativos; elementos que el sistema considera como no válidos en su predicción y que realmente lo son.

- **Sensibilidad** = VP/P = VP/(VP+FN)
- **Especificidad** = VN/N = VN/(FP+VN)

La forma de representación de una curva ROC es la de **Sensibilidad** (tasa de VP) *vs.* **1 - Especificidad** (tasa de FP). Cada punto de la curva ROC representa una tasa comparativa entre ambas medidas en un punto de corte concreto de la población que se analiza. En el tipo de análisis realizado en este trabajo, el algoritmo predictor devuelve un listado con todas las proteínas -o genes- de la red, en orden, desde la que considera como más probable para interactuar con el conjunto inicial (*datos de referencia*) hasta la que menos. La construcción de la curva ROC asociada se lleva a cabo con los siguientes datos: el eje *x* indica la tasa de FP (normalizado en tanto por uno) y el eje *y* en qué porcentaje de ocasiones (en tanto por uno también) encontramos la solución correcta (la proteína/gen del set positivo -P-) en esa posición de la lista o superior (tasa de VP) [141] (véase la figura 3.7 como ejemplo de representación de una curva ROC).

Un predictor perfecto sería aquel cuya curva ROC alcanzase la esquina superior izquierda del área del gráfico (un 100 % de sensibilidad y 100 % de especificidad). Por lo tanto, cuanto más cercana tengamos la curva a la parte superior izquierda, mejor será nuestro predictor. Por el contrario,

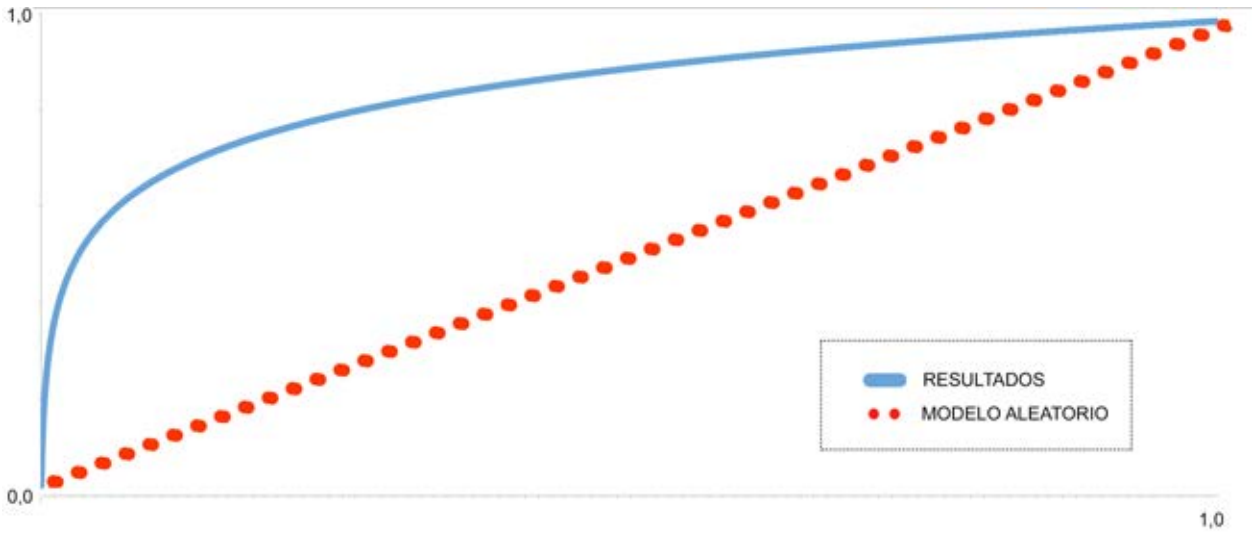


Figura 3.7: **Ejemplo de curva ROC.** El rendimiento obtenido por el predictor de ejemplo está determinado por la curva azul (curva ROC), y más concretamente por el área bajo dicha curva (AUC). La línea roja discontinua representa al modelo aleatorio. Esta curva ROC fue calculada aplicando el método de validación cruzada *Leave One Out* (LOO) sobre el predictor, el cual fue construido para la identificación de nuevas proteínas implicadas en angiogénesis humana (proceso de formación de vasos sanguíneos) [14], tomando como conjunto de datos de referencia a un grupo de proteínas que previamente se conocían implicadas en dicho proceso y aplicando el algoritmo de análisis RWR [127–129] (véase el capítulo 3.2.1) al interactoma. El eje *x* indica la tasa de falsos positivos (FP) en tanto por uno y el eje *y* la cantidad de proteínas, en tanto por uno también, del *set de referencia* que fueron puntuadas en esa posición o superior (tasa de VP).

una distribución aleatoria de las predicciones (sin ningún tipo de valor predictivo) dibujará una línea similar a la diagonal de 45 grados en el espacio de coordenadas: donde las tasas de verdaderos y falsos positivos se igualan para todos los puntos de corte. Por lo tanto, la diagonal divide el espacio ROC: aquellos puntos alejados por encima de la diagonal representan resultados de predicción mejores que los esperados por azar, mientras que puntos cercanos a la diagonal representan malos resultados (similares al azar). La figura 3.8 ilustra lo expuesto anteriormente.

La necesidad de una medida cuantificable a la hora de realizar una comparativa de la calidad de los predictores lleva al concepto del área bajo la curva o **AUC**.

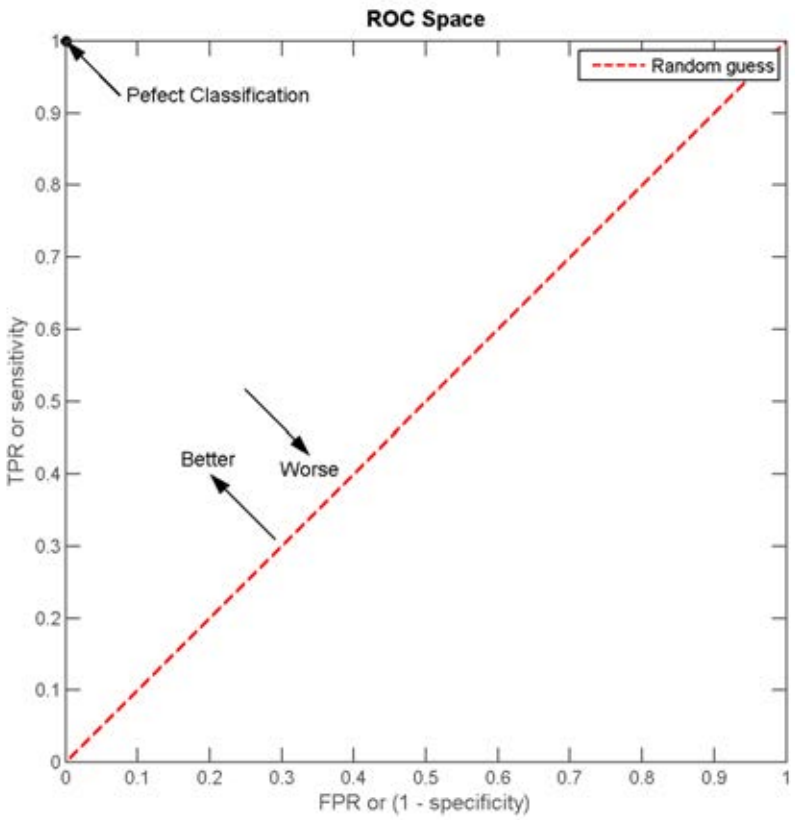


Figura 3.8: **Significado de las diferentes zonas en una curva ROC.** La diagonal representa los valores esperados por azar. Las curvas más próximas a la diagonal (parte inferior derecha) corresponderán a peores predictores que aquellas más alejadas hacia el extremo superior izquierdo. El predictor perfecto es el que alcanza dicha esquina.

El **AUC** (*Area Under the Curve*, o área bajo la curva) es una medida numérica del rendimiento de un sistema predictor, que se basa en el cálculo del área delimitada bajo una curva ROC. Esta área posee un valor comprendido entre **1** (100 %), representando al predictor perfecto (todo el cuadro de coordenadas se encuentra bajo la curva ROC y existe una especificidad y sensibilidad totales) y **0,5** (50 %), asociado a un rendimiento similar al esperado por azar y sin capacidad predictiva alguna (solamente la mitad inferior derecha se encuentra bajo la curva -representación diagonal-). Véase la figura 3.9.

Este valor se puede interpretar como la probabilidad de que, tomados al azar un caso positivo y uno negativo, la puntuación obtenida para el primero sea superior a la del segundo. Es decir, si el AUC para un predictor, de los desarrollados en esta Tesis Doctoral, es 0,7 significa que existe un 70 % de probabilidad de que la puntuación del predictor para una proteína positiva (que sí pertenece al sistema molecular) sea mayor que para una negativa (que no pertenece al sistema molecular), seleccionadas aleatoriamente.

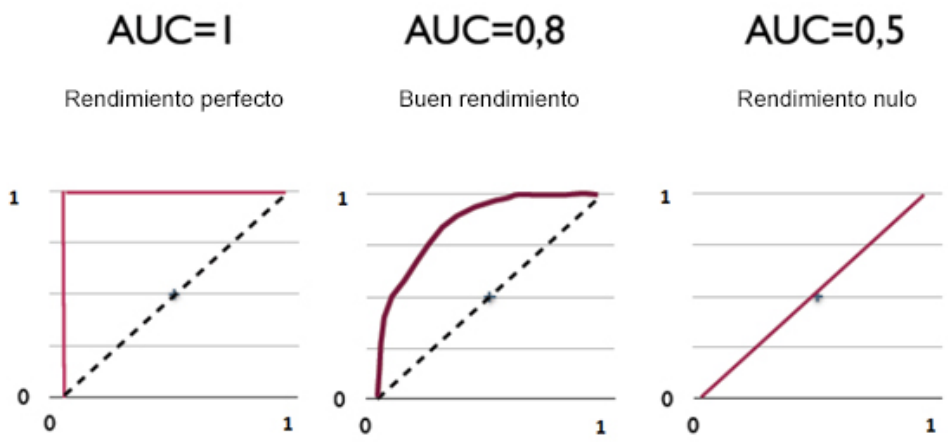


Figura 3.9: **Ejemplos de valores del AUC y sus respectivas curvas ROC.**

Para la elección entre sistemas predictores se recurre a la comparativa de las curvas ROC y sus valores asociados del AUC, eligiendo al que presente un mayor AUC, es decir una mejor capacidad predictiva.

Como guía interpretativa se han establecido los siguientes intervalos para valores de AUC:

- 0,5: Igual al azar.
- 0,5 - 0,6: Mal rendimiento.
- 0,6 - 0,75: Rendimiento intermedio.
- 0,75 - 0,9: Buen rendimiento.
- 0,9 - 0,99: Rendimiento excelente.
- 1: Rendimiento perfecto.

3.4. Redes biomédicas para la identificación de relaciones genotipo-fenotipo en enfermedades raras

En el estudio detallado en el capítulo 6 se analizó la información biomédica de la base de datos DECIPHER [62] (previa firma de un convenio con el consorcio de la plataforma), y a través de la información anónima allí contenida sobre: fenotipos, genotipos y pacientes de enfermedades raras se creó una red heterogénea que interconectaba estas 3 capas de información para su posterior análisis mediante las técnicas algorítmicas detalladas en el capítulo 3.2.2, más concretamente a través del algoritmo de análisis de redes heterogéneas: *Índice Hipergeométrico*. La finalidad fue la de identificar relaciones genotipo-fenotipo estadísticamente significativas en este tipo de patologías. Este procedimiento dio lugar a un sistema predictor de relaciones CNV-fenotipo (véase el capítulo 6).

Cabe destacar que las anotaciones fenotípicas de los pacientes (procedentes de DECIPHER [62]) se llevaron a cabo haciendo uso de *The Human Phenotype Ontology* (HPO) [10, 52, 53], teniendo así un vocabulario estándar, sin ambigüedades y correctamente jerarquizado para su uso sistemático.

Los detalles metodológicos concretos, así como los resultados de este estudio, se pueden consultar en el capítulo 6 y la publicación asociada, presentada como aval de esta Tesis Doctoral, en el capítulo 6.3.4.

3.5. Lenguajes de programación y medios computacionales empleados

Para el desarrollo de los estudios incluidos en esta Tesis Doctoral se han utilizado diferentes lenguajes de programación informática. La justificación del uso de cada uno de ellos es como sigue: se ha utilizado *Python* para el preprocesado de datos, el trabajo con ficheros de texto plano y el resto de cálculos sencillos. Esta elección se debe a que se trata de un lenguaje interpretado multiplataforma, por lo que no requiere compilación y permite un sencillo y ubicuo manejo de ficheros. Para los cálculos complejos (*e.g.* CT, DK o los análisis de redes *tripartitas*) y la generación de modelos aleatorios (lo cual requiere de alta capacidad computacional) se ha utilizado lenguaje *MATLAB*, por presentar un alto rendimiento en este entorno y ser también la forma más sencilla de implementarlos, a través de un lenguaje orientado a la formulación matemática y la posibilidad de paralelización en *Picasso* (supercomputador alojado en la Universidad de Málaga -figura 3.10-). Con respecto a los cálculos estadísticos y la generación de gráficas, se llevaron a cabo en entorno *R* [47], de manera local, por ser una herramienta muy potente para problemas estadísticos con grandes cantidades de datos y su representación gráfica. Véase la figura 3.11.

Figura 3.10: **Supercomputador Picasso (Centro de Supercomputación y Bioinformática - Red Española de Supercomputación. Universidad de Málaga).**

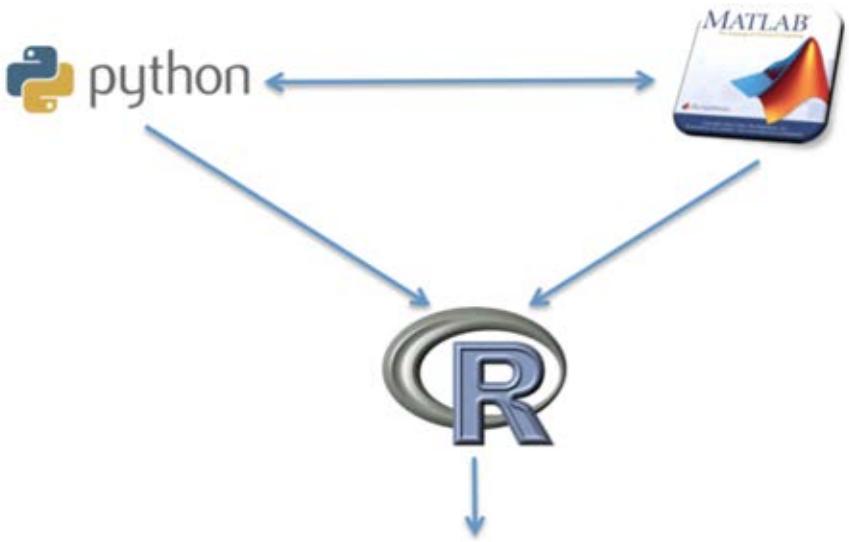


Figura 3.11: **Entornos de programación utilizados en el estudio.**

Es de vital importancia cuando se trata con grandes cantidades de datos, como es el caso de este trabajo, optimizar el código producido. En este caso se ha llevado a cabo un proceso de rediseño progresivo de la estructura de los algoritmos según los requerimientos de cada situación. Se hicieron ejecuciones de prueba con tal de comprobar la eficiencia del algoritmo en cada punto. En general, se ha optimizado el código según los siguientes criterios:

- Reservando memoria de cómputo para las grandes matrices únicamente cuando era necesario y liberándola en el momento en el que ya no lo era.
- Generando estructuras de datos de tamaño mínimo.
- Eliminando código innecesario.
- Extrayendo código de los bucles cuando era posible.
- Pasando objetos por referencia en lugar de por valor.
- Minimizando y optimizando el acceso a disco.

En lo relativo a la ejecución, se llevó a cabo de manera local, excepto en los casos en los que la capacidad de cómputo y la paralelización eran cruciales: el cálculo de los resultados de DK y CT, y la generación de los modelos aleatorios. En estos casos se hizo uso del supercomputador Picasso en el SCBI (Centro de Supercomputación y Bioinformática) de Andalucía, dentro de la Red Española de Supercomputación y con base en la Universidad de Málaga (véase la figura 3.10). Este supercomputador tenía una arquitectura PPC64 (IBM PowerPC 970) con 512 procesadores, un rendimiento de 2994,04 GFlops y 1024 Gb de memoria; en el momento de desarrollar este trabajo.

Capítulo 4

Implementación de métodos de predicción funcional basados en redes de interacción proteína-proteína: aplicación a sistemas de diferenciación maligna de células tumorales y a la angiogénesis

El aspecto más triste de la vida ahora mismo es que la Ciencia aporta conocimiento más rápido de lo que la sociedad desarrolla sabiduría.

Isaac Asimov

4.1. Antecedentes

El desarrollo de predictores funcionales de proteínas que aquí se presenta fue iniciado tras una propuesta de colaboración por parte del grupo de investigación *Cell biology of cancer* del *Karolinska Institute* (Estocolmo), dirigido por el Dr. Staffan Strömblad. Dicho grupo había realizado trabajos previos de transcriptómica y proteómica sobre células tumorales de un determinado tipo de cáncer de mama en 2 estados diferentes: reposo y fenotipo maligno. El cambio al fenotipo maligno se veía relacionado con la rigidez de la matriz extracelular.

Se llevó a cabo la construcción de unas herramientas bioinformáticas predictivas con la finalidad de identificar qué proteínas de las expresadas diferencialmente de manera significativa entre ambos estados podrían estar implicadas en los procesos moleculares que detectaban, transmitían y ejecutaban el cambio a un fenotipo maligno en dichas células en respuesta a esa modificación de la rigidez en el medio.

Tras completar los predictores y validarlos con las técnicas que se detallan en este capítulo, por problemas personales, el investigador principal en el desarrollo experimental de la colaboración por parte del *Karolinska Institute* abandonó el proyecto. No obstante, se decidió aprovechar la metodología ya desarrollada y adaptarla a otro proyecto, en el ámbito de la angiogénesis, que había en marcha en el Departamento de Biología Molecular y Bioquímica de la Universidad de Málaga, dando fruto a la siguiente publicación que se adjunta al final de este capítulo: **García-Vilas JA, Morilla I, Bueno A, Martínez-Poveda B, Medina MÁ, Ranea JAG.** ***In silico prediction of targets for anti-angiogenesis and their in vitro evaluation confirm the involvement of SOD3 in angiogenesis.*** **Oncotarget 2018; 9.** (capítulo 5.3.6). En dicho trabajo se adaptó y aplicó la metodología que se detalla a continuación para predecir nuevas proteínas implicadas en angiogénesis, y posteriormente se llevó a cabo una validación experimental de alguna de esas predicciones.

Es por todo lo anterior que a lo largo de este capítulo se detalla una metodología en base a un supuesto biológico (sobre el cual se desarrolló inicialmente el sistema predictor y su validación

teórico-matemática), como es el cambio de células tumorales a estado maligno en situación de rigidez en la matriz extracelular; y finalmente se adjunta una publicación como resultado de su aplicación en otro contexto biológico, como es el proceso de angiogénesis, en el cual se validó experimentalmente con éxito dicha metodología predictiva.

4.2. Introducción

4.2.1. Diferenciación maligna de células tumorales inducidas por la rigidez de la matriz extracelular

Se ha observado experimentalmente que células tumorales implicadas en cáncer de mama (de la línea MCF10CA1a) cambian hacia un fenotipo maligno [11] cuando el sustrato donde se asientan incrementa su rigidez [12, 13]. En base a 2 estados celulares: i) células tumorales en reposo *vs.* ii) células tumorales con fenotipo maligno inducido por un sustrato más rígido; se llevaron a cabo (por parte del *Karolinska Institute*) experimentos de transcriptómica y proteómica para determinar los genes que variaban su expresión significativamente -a nivel de ARN mensajero y de la cantidad de proteína producida- para dichos estados: de alto grado de rigidez en la matriz extracelular (5000 Pa) a bajo (400 Pa). Estas aproximaciones experimentales arrojaron un alto número de genes con expresión diferencial significativa pero con implicaciones funcionales muy diversas y en algunos casos desconocidas. Se construyó un sistema predictor mediante una aproximación sistémica, haciendo uso de las fuentes de datos descritas en el capítulo 3.2 de Materiales y métodos generales (redes de interacción de proteínas) y las metodologías expuestas en el capítulo 3.2.1 (algoritmos de análisis de redes DK y CT), con el objetivo de predecir qué genes de los que muestran expresión diferencial con el cambio de rigidez podrían potencialmente estar implicados en procesos oncogénicos (de transformación a un fenotipo maligno): transducción de señales mecanosensoras, unión de la célula al sustrato, vías de señalización oncogénicas, etc.

El primer paso fue identificar y definir los sistemas moleculares conocidos dentro del mecanosensor que se sabían implicados tanto en la detección como en la transmisión de señal asociada a cambios en la rigidez de la matriz extracelular (como se ha visto, relacionado con el cambio de malignidad en las células tumorales); para ello se acudió a conocimiento experto, bibliografía disponible [13, 16] y bases de datos: de interpretación funcional (DAVID [17, 18]), de rutas metabólicas (KEGG [19]) y de procesos celulares (GO [20]). Los sistemas finalmente identificados

pueden verse en la figura 4.1 y son los siguientes: i) adherencia célula-célula, ii) regulación del citoesqueleto, iii) adhesión focal, iv) ruta de señalización Hippo, v) regulación de la mecánica nuclear y vi) mecanotransducción de la señalización oncogénica. En segundo lugar se creó una lista, para cada uno de estos sistemas, incluyendo a todos los genes/proteínas que se conocían implicados en los mismos. Estas listas constituyen los grupos o *sets de referencia*.

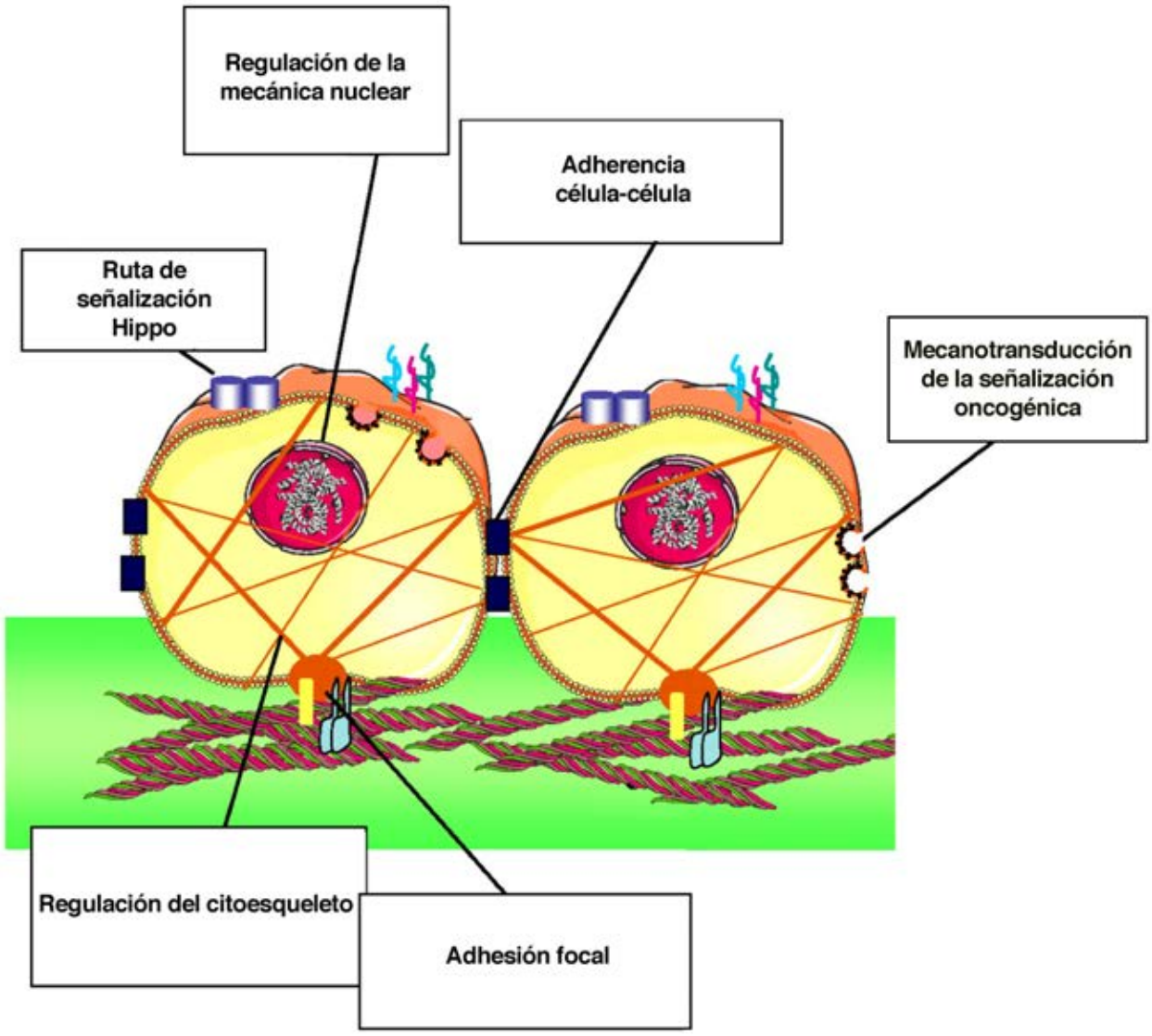


Figura 4.1: **Sistemas moleculares implicados en el cambio hacia un fenotipo maligno de las células tumorales MCF10CA1a.**

4.2.2. Implementación de los predictores basados en redes

Una vez establecido el *set de referencia* para cada uno de los sistemas, se hizo uso de la información disponible en las bases de datos de redes de interacción de proteínas, correspondiente a todo el proteoma humano (véase el capítulo 1.3 de la Introducción general) y se les aplicó un análisis mediante técnicas *kernel* con la finalidad de calcular las distancias funcionales entre los contextos de interacciones de cada par de proteínas del interactoma (véase el capítulo 3.2.1 de Materiales y métodos generales) para así poder valorar numéricamente la probabilidad de que cada uno de los genes/proteínas del conjunto *problema* (el proteoma humano) estuviese implicado en cada uno de los sistemas moleculares antes descritos. Dichos predictores de genes candidatos a estar implicados en estos sistemas celulares se validaron estadísticamente aplicando la metodología *Leave One Out* -LOO- y curvas *Receiver Operating Characteristic* -ROC- (véase el capítulo 1.3.2 de la Introducción general) con la finalidad de dilucidar si esa distancia en red permitía predecir casos positivos para cada sistema molecular.

Este procedimiento computacional sistemático ya ha demostrado ser eficiente realizando predicciones a la hora de seleccionar, en genomas completos, pequeños grupos de genes candidatos asociados a sistemas biológicos concretos [46, 142–145].

Si una proteína *problema* se encuentra *cerca* en la red de interacciones (más de lo esperado por azar) del conjunto de proteínas de un sistema implicado en la diferenciación celular mencionada, podemos inferir que dicha proteína juega, probablemente, un papel funcional en dicho sistema molecular. Esto podría hacerla seleccionable para llevar a cabo validaciones experimentales.

Se hizo uso de distintos modelos de redes de interacción de proteínas (modelos del interactoma humano) y de distintas métricas de distancias en red de tipo *kernel* para desarrollar una batería de predictores funcionales específicos para cada uno de los sistemas celulares del proceso mecanosensor antes detallados. En cada uno de los sistemas se implementó la metodología con todos los modelos de red disponibles y los diferentes algoritmos de análisis *kernel*, es decir, se proba-

ron y compararon todas las combinaciones posibles de métrica y red. Posteriormente se evaluó estadísticamente el rendimiento de cada combinación de modelo del interactoma y métrica usada, mediante la técnica LOO y curvas ROC, con el fin de definir aquellos con mejores resultados (véase el capítulo 1.3.2 de la Introducción general).

Estos predictores generan una lista priorizada de genes/proteínas candidatos ordenados por su *cercanía* estadística en la red al *set de referencia* [139], es decir, a los distintos subprocesos moleculares específicos que median en la diferenciación maligna de células tumorales MCF10CA1a, en respuesta al cambio en la rigidez del sustrato. Cabe destacar que el método evita además los defectos en la métrica derivados de artefactos locales en la topología de la red, tales como nodos excesivamente conectados (proteínas promiscuas). Véase el capítulo 3.2.1 de Materiales y métodos generales para más información.

Para cada uno de los 6 sistemas moleculares antes descritos, asociados al proceso oncogénico y de adhesión celular, se seleccionó la combinación de métrica y red que mostró el mejor rendimiento predictivo, haciendo uso de las evaluaciones de rendimientos con las técnicas ya mencionadas (tal como también se hizo en García-Vilas *et al.* [14], véase el capítulo 4.4.3). Para más información acerca de la metodología de validación se puede consultar el capítulo 1.3.2 de la Introducción general y el capítulo 3.3 de Materiales y métodos generales.

Finalmente, los predictores que mostraron el mejor rendimiento en cada uno de los sistemas fueron utilizados para determinar qué genes, de los que habían mostrado expresión diferencial significativa (genes/proteínas *problema*), se encontraban altamente asociados a dichos sistemas moleculares (a sus *sets de referencia*) a través de las redes de interacción, y por tanto, presentaban una mayor probabilidad de estar directamente implicados en los procesos mecanosensores que conducen a la transformación maligna de las células tumorales estudiadas en casos de cambio en la rigidez de la matriz extracelular.

4.3. Material y métodos

La figura 4.2 resume el proceso general que se ha seguido en el presente estudio para llevar a cabo las predicciones computacionales de genes/proteínas candidatos: los pasos de selección y procesamiento de bases de datos de interacciones de proteínas (modelos del interactoma humano), la construcción de las redes biológicas, la selección de los algoritmos y la medición de distancias de contextos funcionales (*kernel*), la obtención de las puntuaciones y su significancia, la validación del método mediante LOO y curva ROC y la obtención final de las predicciones con su grado de fiabilidad asociado. Los algoritmos aquí implementados han sido descritos y utilizados con éxito en otros trabajos [14, 46, 146].

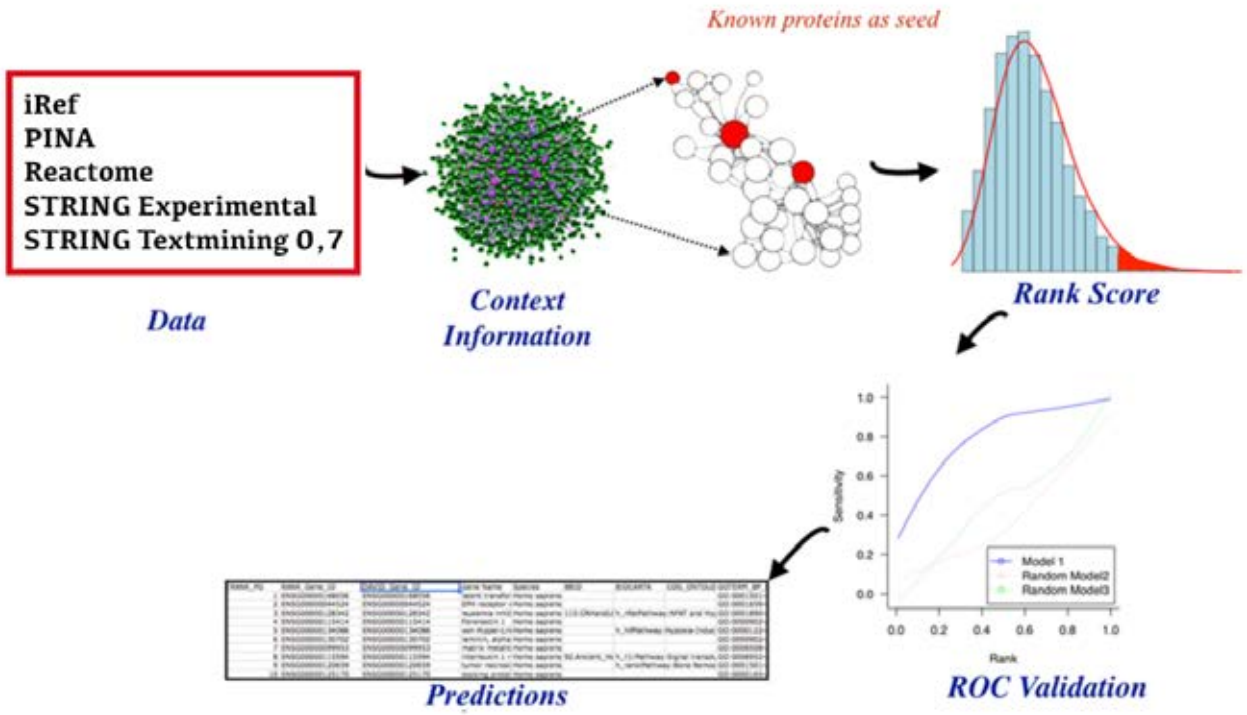


Figura 4.2: **Esquema de la metodología de análisis bioinformática llevada a cabo para la predicción de nuevos genes implicados en procesos celulares.** Adaptado de García-Vilas *et al.* [14].

4.3.1. Obtención de los *sets de referencia*

En base a la bibliografía [13, 15, 16], las bases de datos públicas: de interpretación funcional (DAVID [17, 18]), de rutas metabólicas (KEGG [19]) y de procesos celulares (GO [20]); y las revisiones manuales de expertos, se identificó un conjunto de proteínas con un rol funcional en cada uno de los sistemas dentro de los procesos mecanosensores. Estos sistemas celulares fueron: adherencia célula-célula, regulación del citoesqueleto, adhesión focal, ruta de señalización Hippo, regulación mecánica del núcleo y mecanotransducción de la señalización oncogénica (véase la figura 4.1). Cada uno de los sistemas estaba compuesto por el número de proteínas indicado en la tabla 4.1 como *set de referencia*.

Sistema molecular	# proteínas
Adherencia célula-célula	30
Regulación del citoesqueleto	43
Adhesión focal	77
Ruta de señalización Hippo	9
Regulación de la mecánica nuclear	15
Mecanotransducción de la señalización oncogénica	16

Tabla 4.1: **Número de proteínas que formaron parte del *set de referencia* en cada sistema molecular descrito.**

4.3.2. Bases de datos de interacciones entre proteínas (PPI) y métodos de análisis

5 modelos diferentes del interactoma humano, en la forma de redes de interacción de proteínas, fueron analizados: iRef [21] (que proporciona un índice de interacciones entre proteínas a partir de la consulta a varias bases de datos), PINA [22] (que integra información de 6 bases de datos diferentes, creando un índice completo y no redundante), STRING Experimental (que incluye exclusivamente interacciones validadas experimentalmente) [23], STRING Textmining [23] (con datos derivados de la aparición conjunta de nombres -co-citas- de genes/proteínas en los resúmenes de publicaciones científicas) y Reactome Pathways [24] (con información obtenida de rutas metabólicas y reacciones). En el caso de STRING Textmining, y debido a que esta base de datos dota de un porcentaje de fiabilidad a cada una de las interacciones que contiene, se aplicó un punto de corte al 70 % de fiabilidad, tal como se recomienda en la documentación de la propia base de datos, con la finalidad de trabajar con datos más fiables. Véase el capítulo 3.2 de Materiales y métodos generales para más detalles sobre estas fuentes de datos.

En lo relativo al análisis de asociaciones dentro de las redes para medir la relación -o cercanía- entre las proteínas *problema* y los respectivos *sets de referencia* (proteínas que sí se sabe que están implicadas en los diferentes sistemas moleculares), se descartaron los métodos de distancia directa (por las razones ya comentadas en el capítulo 3.2.1 de Materiales y métodos generales), y se aplicaron, a las matrices que representan las redes, 2 algoritmos diferentes basados en probabilidades de difusión (*kernel*): **Commute Time Kernel** (CT) [$K = L^{+}$] y **Exponential Laplacian Diffusion Kernel** (DK) [K = exp(-βL)] debido a que son los algoritmos que han mostrado, en estudios recientes [46], un buen rendimiento en el análisis adecuado de las relaciones contextuales en red de sistemas biológicos, ya que, entre otras cosas, tienen en cuenta la topología del grafo de interacciones, tal como se ha detallado en la correspondiente sección de Materiales y métodos generales (capítulo 3.2.1).

El uso de varias redes de interacción de proteínas y de 2 algoritmos de análisis de estas redes dota de un marco comparativo con el que poder seleccionar las combinaciones con mejor rendimiento.

A través de estos diferentes modelos de redes de interacción proteína-proteína y de las métricas de asociación empleadas, se construyó y validó un conjunto de predictores específico para cada uno de los sistemas moleculares mencionados anteriormente. Esto es, para cada uno de los 6 sistemas se generaron **10** predictores: utilizando 5 redes de interacción distintas (modelos del interactoma) y 2 métodos de cálculo de distancias, es decir, 10 combinaciones diferentes (véase la figura 4.3).

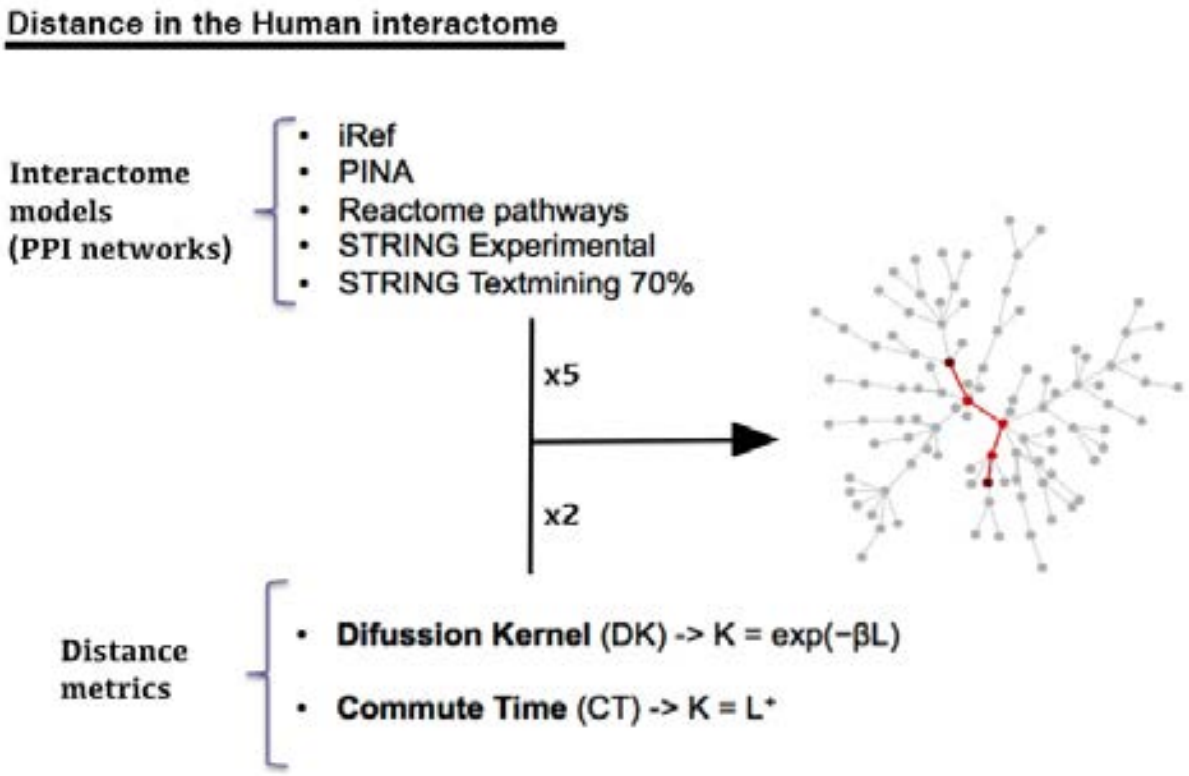


Figura 4.3: **Bases de datos y algoritmos utilizados para el análisis predictivo:** todas las bases de datos usadas como modelos de interacción de proteínas y los diferentes tipos de métricas propuestas.

4.3.3. Creación y validación de los predictores

De entre los 10 predictores construidos, para cada uno de los 6 sistemas moleculares, se seleccionó la combinación de métrica y modelo de red que mostró el mejor rendimiento (es decir, 1 de esos 10 predictores). Para medir dichos rendimientos se hizo uso del método de validación cruzada LOO y el test ROC, junto con el cálculo del AUC (*Area Under the Curve*) asociado. Estos métodos ya fueron utilizados de manera exitosa previamente en estudios similares [46, 146] y han sido detallados en el capítulo 3.3, Materiales y métodos generales.

Más concretamente, el proceso siguió los siguientes pasos:

Fase I - Construcción de los predictores

- Se calculó, haciendo uso de las 2 métricas descritas, la distancia entre todos los pares de proteínas en cada una de las redes de interacción (modelos del interactoma), obteniendo las 10 matrices de distancias de tamaño *n x n* (siendo *n* el número de proteínas en cada modelo del interactoma).

- Para cada uno de los 6 sistemas moleculares (véase la figura 4.1), teniendo los datos de distancias en red del punto anterior, se calculó, para cada red PPI y métrica, la *distancia* funcional de cada proteína del interactoma con respecto a las proteínas consideradas como *set de referencia*, para ello se aplicó una media aritmética de las distancias individuales de la proteína *problema* a cada una de las proteínas pertenecientes a este *set* (conjunto de proteínas que sí se saben implicadas en el sistema molecular). Véase la figura 4.4.

- Se construyó una lista ordenada de las proteínas de todo el interactoma en función de su *cercanía* funcional (medida a través de la media aritmética del punto anterior) al *set de referencia* en cada caso, de mayor *cercanía* a menor. De este modo ya se dispone de los resultados de los predictores.

Una vez construidos los predictores se llevó a cabo la validación LOO/ROC con la finalidad de seleccionar, para cada sistema molecular, la red y método de análisis que daba mejores resultados (de entre los 10 disponibles).

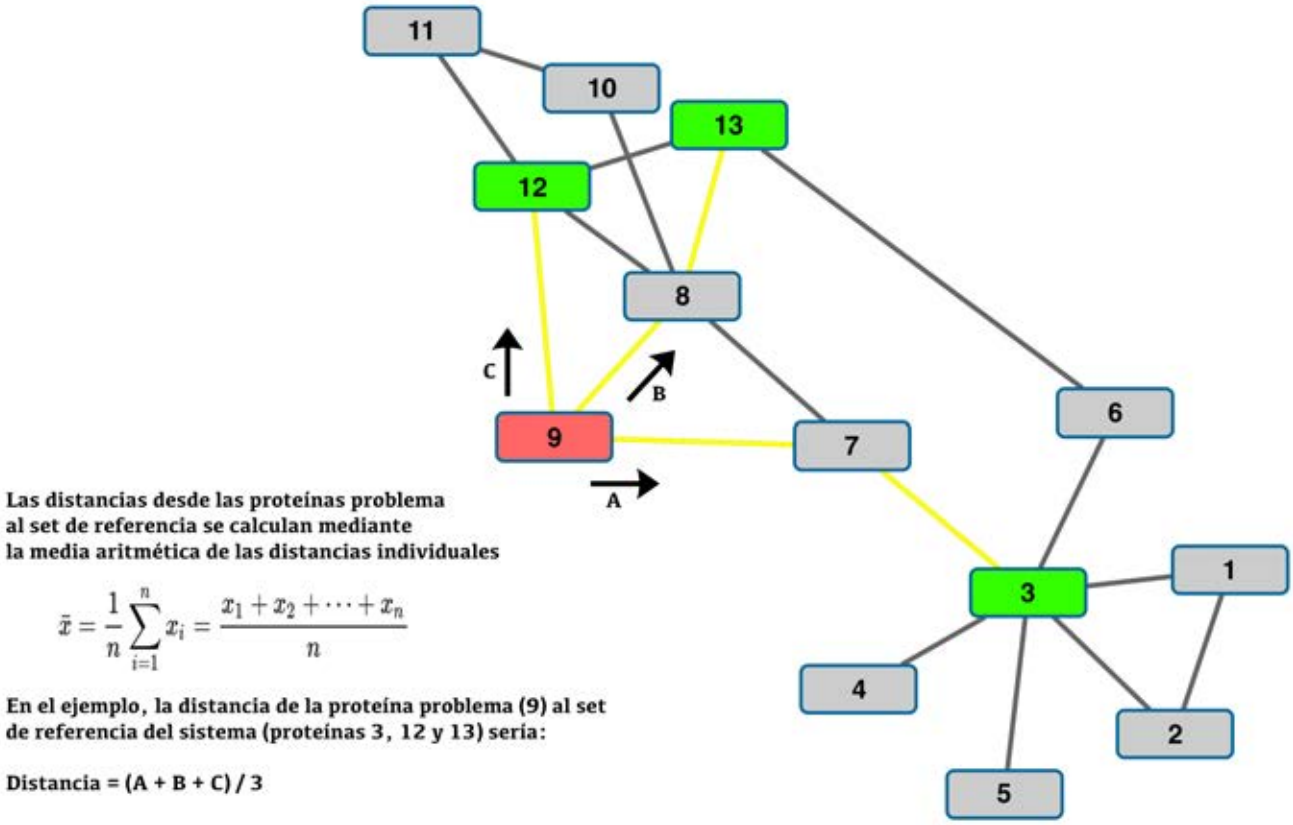


Figura 4.4: **Ejemplo de cálculo de distancia de una proteína *problema* al *set de referencia* asociado a un proceso molecular estudiado.** La distancia de una proteína con respecto a un conjunto de proteínas de referencia de un proceso molecular se calcula, haciendo uso de las redes de interacción entre proteínas, con una media aritmética de las distancias de dicha proteína a cada una de las que conforman el *set de referencia*, de manera individual. En este caso, teniendo la proteína *9* como proteína *problema* (en rojo) y las proteínas *3*, *12* y *13* como *set de referencia* (en verde), se calcula primero la distancia entre *9* y *3* **(A)**, entre *9* y *13* **(B)** y entre *9* y *12* **(C)**; y posteriormente se hace un media aritmética de ellas. Cabe destacar que aquí se han representado las *distancias* mediante caminos de aristas amarillas (método de cálculo de la distancia más corta) para simplificar la abstracción de las técnicas *kernel* aplicadas en este estudio, pero el procedimiento de obtención de la distancia global (el cálculo de la media aritmética entre las *distancias*) es el mismo.

Fase II - Validación estadística de los predictores: LOO y curvas ROC

- Se realizó un proceso *Leave One Out* (LOO) sobre los predictores, repitiendo los cálculos tantas veces como número de proteínas hubiese en los *set de referencia*. Es decir, en cada iteración se 'desanotó' del *set de referencia* una de las proteínas que pertenecía al mismo, se realizaron los cálculos sobre la red (teniendo ahora un *set de referencia* de *n* - 1 proteínas, siendo *n* el número inicial) y se observó en qué posición del *ranking* de predicciones aparecía esa proteína 'desanotada' que había pasado a ser parte de las proteínas *problema*. Véase el capítulo 3.3.1 de Materiales y métodos generales para más detalles.

- Se recopilaron los datos de todas las posiciones que adquirían en el listado las propias proteínas del *set de referencia* cuando se las consideraba como proteínas *problema* (casos positivos). Se hizo lo mismo con un conjunto aleatorio de proteínas con el mismo tamaño que el *set de referencia* (casos negativos). Y con estos datos se construyeron curvas ROC. El eje *x* de la gráfica indica la **tasa de falsos positivos**, que en este caso se corresponde con la posición en el *ranking* -punto de corte-, pues en el modelo aleatorio los positivos se distribuyen de manera uniforme en la lista priorizada. Los valores se expresan en tanto por uno y comenzando por las posiciones con mejor puntuación. El eje *y* indica la **tasa de verdaderos positivos**, es decir la cantidad de proteínas (en tanto por uno también) del *set de referencia* que fueron puntuadas en esa posición o superior -punto de corte- cuando fueron 'desanotadas' y consideradas como proteínas *problema*. Cuanto mayor es el área que queda bajo la curva mejor es el sistema predictor (mayor número de proteínas pertenecientes al sistema -*set de referencia*- correctamente catalogadas entre las primeras posiciones). Esta aproximación gráfica se cuantifica también mediante lo que se conoce como AUC (*Area Under the Curve*), cuyos valores van de 1 (predictor perfecto) a 0,5 (predictor aleatorio). Como modelo aleatorio se toma la línea diagonal, ya que por azar una proteína puede ser catalogada en cualquiera de las posiciones de la lista de priorización (en cualquiera de los valores otorgados a proteínas del interactoma). Véase el capítulo 3.3.2 de Materiales y métodos generales.

- Una vez construidas las curvas ROC mediante el método LOO se pudo valorar visualmente (mediante las gráficas) y numéricamente (mediante el cálculo del AUC -*Area Under the Curve*-) el rendimiento (véase el capítulo 3.3.2) de cada uno de los predictores (compuestos por una red de interacciones concreta y un algoritmo de análisis concreto) y de este modo determinar para cada sistema molecular cuál de ellos mostraba un mejor rendimiento.

Fase III - Obtención de las listas de priorización

- Tras la construcción de los predictores, la validación de los mismos mediante LOO y ROC y la selección del más apropiado para cada sistema molecular asociado a la diferenciación celular maligna debida al cambio de la rigidez de la matriz extracelular en casos de células cancerosas de la línea MCF10CA1a; ya se tiene un listado priorizado (haciendo uso del predictor seleccionado) de todas las proteínas del interactoma en relación a cada *set de referencia* (cada sistema molecular).

4.4. Resultados y Discusión

El proceso biocomputacional, resumido en la figura 4.2, incluyó: obtención de datos de interacciones de proteínas, extracción de información basada en el contexto funcional mediante análisis a través de métricas de redes (teniendo como referencia las proteínas pertenecientes a cada sistema molecular), validación mediante LOO y curva ROC y obtención del listado final puntuado y ordenado de proteínas candidatas a estar implicadas en el sistema molecular estudiado.

4.4.1. Curvas ROC para cada uno de los sistemas moleculares

Los resultados de las curvas ROC para cada uno de los sistemas moleculares estudiados (véase la figura 4.1) se muestran en las páginas siguientes (figuras de la 4.5 a la 4.10) y permiten evaluar visualmente, en cada caso, cuál de los predictores (combinación de red PPI y algoritmo utilizado) tiene, *a priori*, un mejor rendimiento.

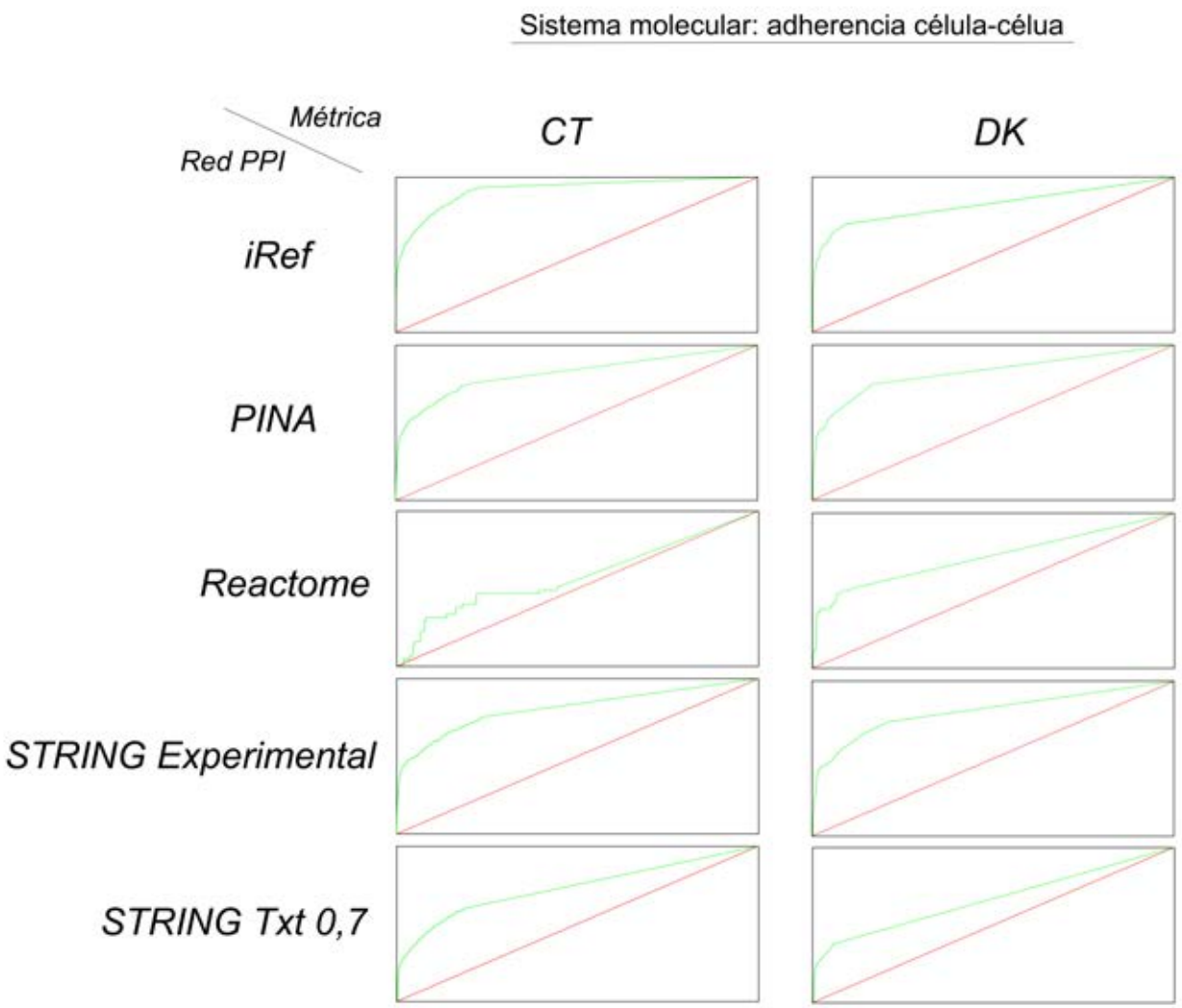


Figura 4.5: **Curvas ROC de los predictores para el sistema *adherencia célula-célula*.** Se muestran los resultados de las curvas ROC haciendo uso de diferentes redes de interacción de proteínas y métricas de análisis. El eje *x* de la gráfica indica la tasa de falsos positivos (en tanto por uno) y el eje *y* la cantidad de proteínas (en tanto por uno también) del *set de referencia* que fueron puntuadas en esa posición o superior (tasa de verdaderos positivos). La diagonal, en rojo, representa el modelo aleatorio.

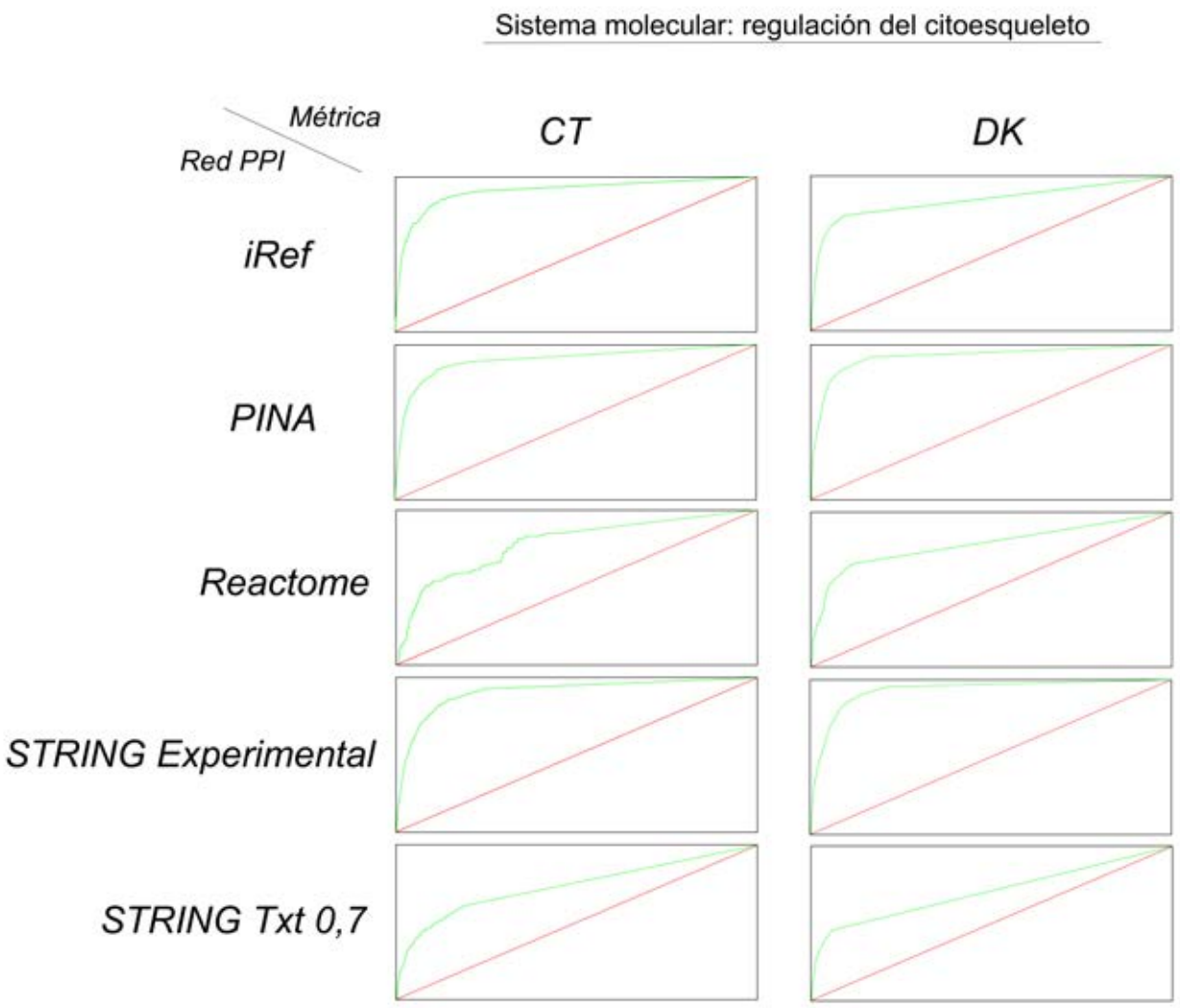


Figura 4.6: **Curvas ROC de los predictores para el sistema *regulación del citoesqueleto*.** Misma descripción que en la figura 4.5.

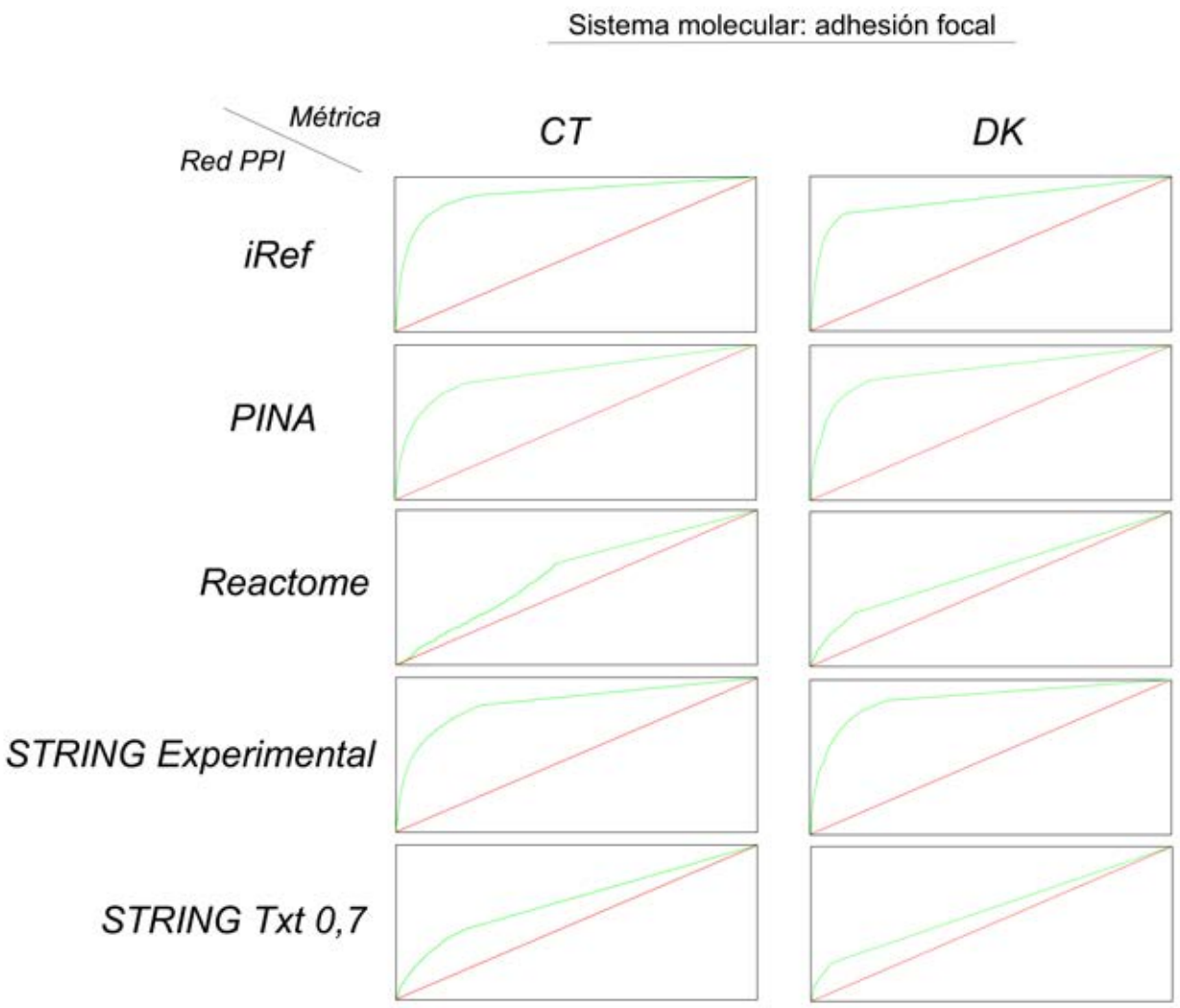


Figura 4.7: **Curvas ROC de los predictores para el sistema *adhesión focal*.** Misma descripción que en la figura 4.5.

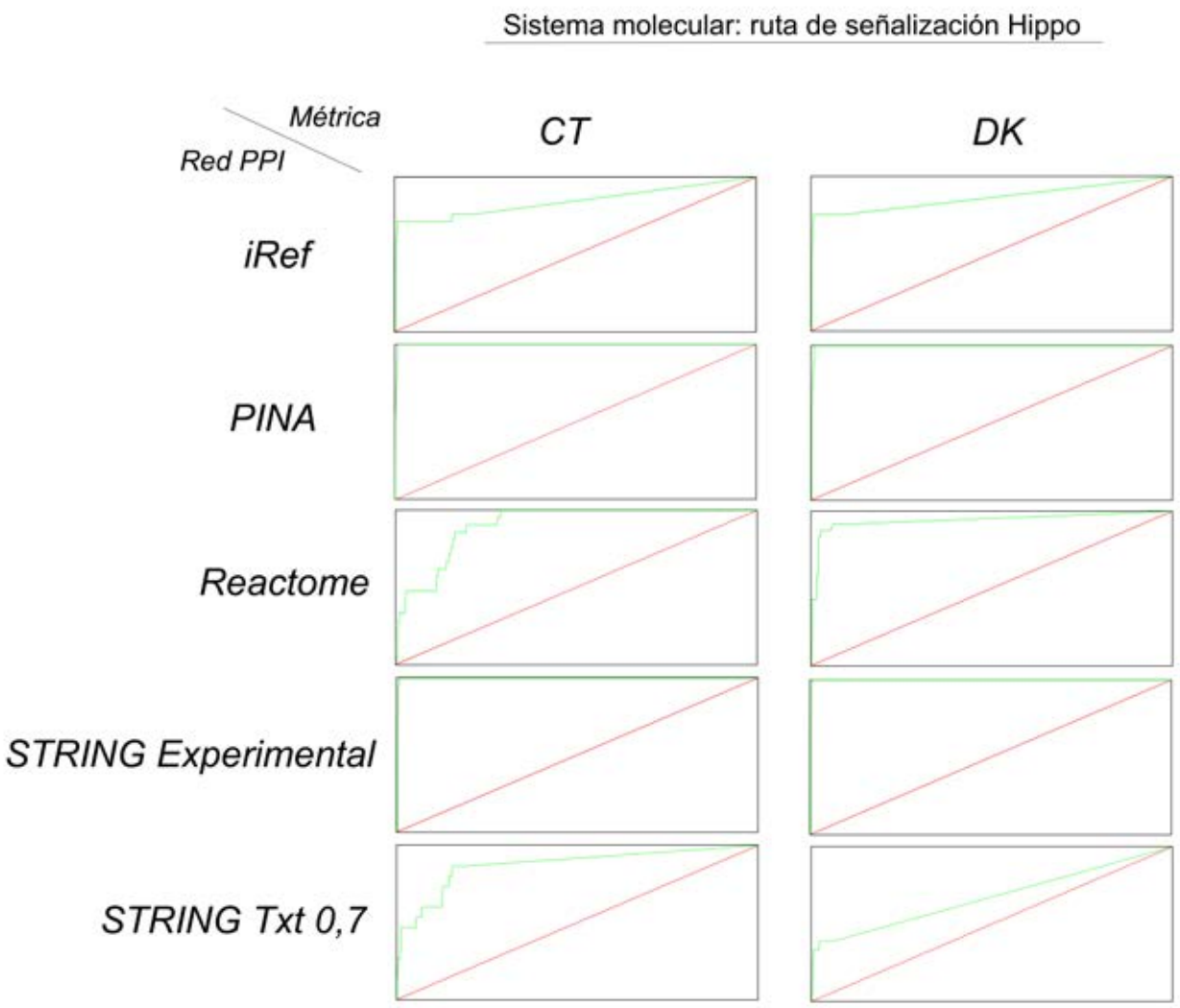


Figura 4.8: **Curvas ROC de los predictores para el sistema *ruta de señalización Hippo*.** Misma descripción que en la figura 4.5.

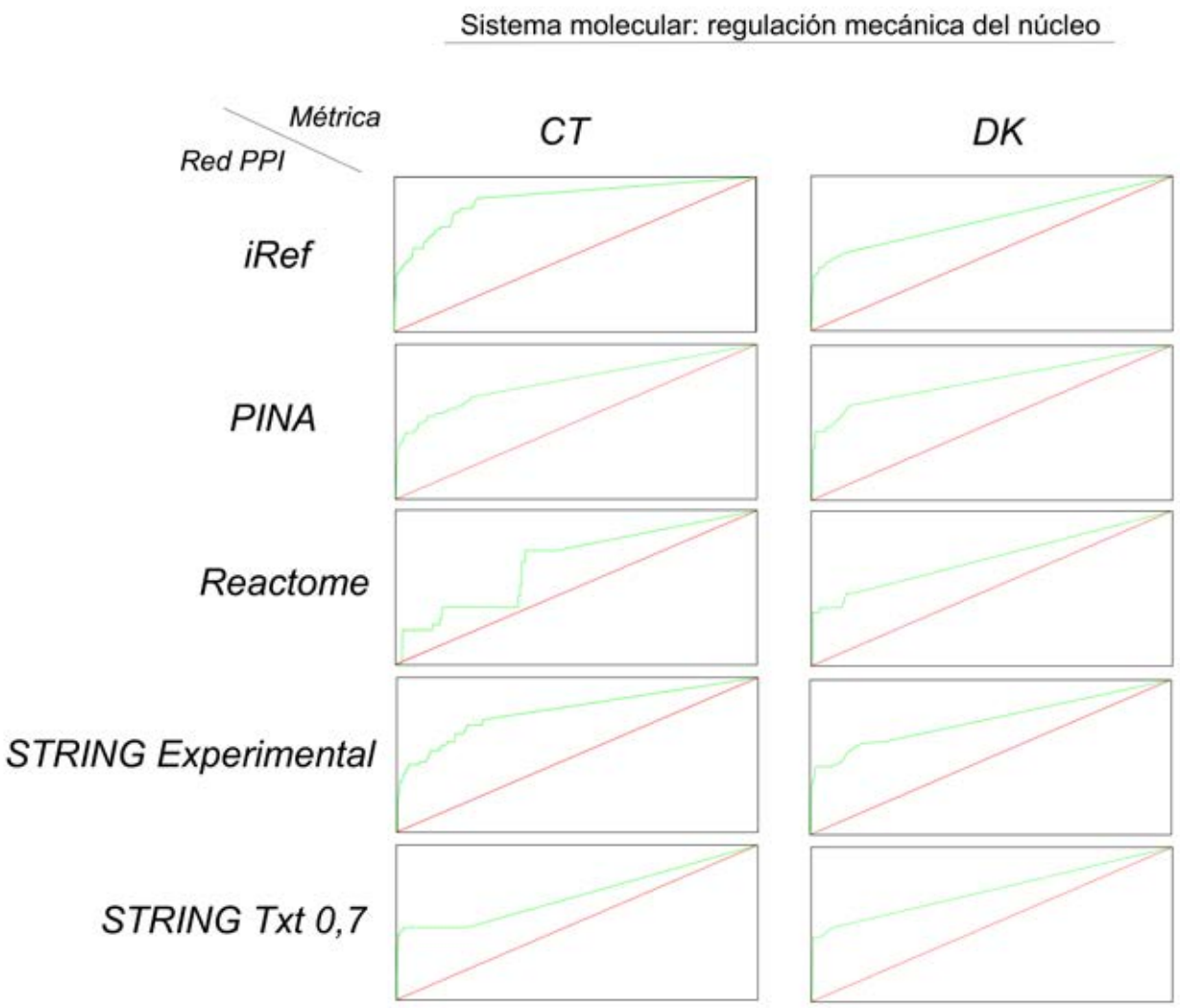


Figura 4.9: **Curvas ROC de los predictores para el sistema *regulación mecánica del núcleo*.** Misma descripción que en la figura 4.5.

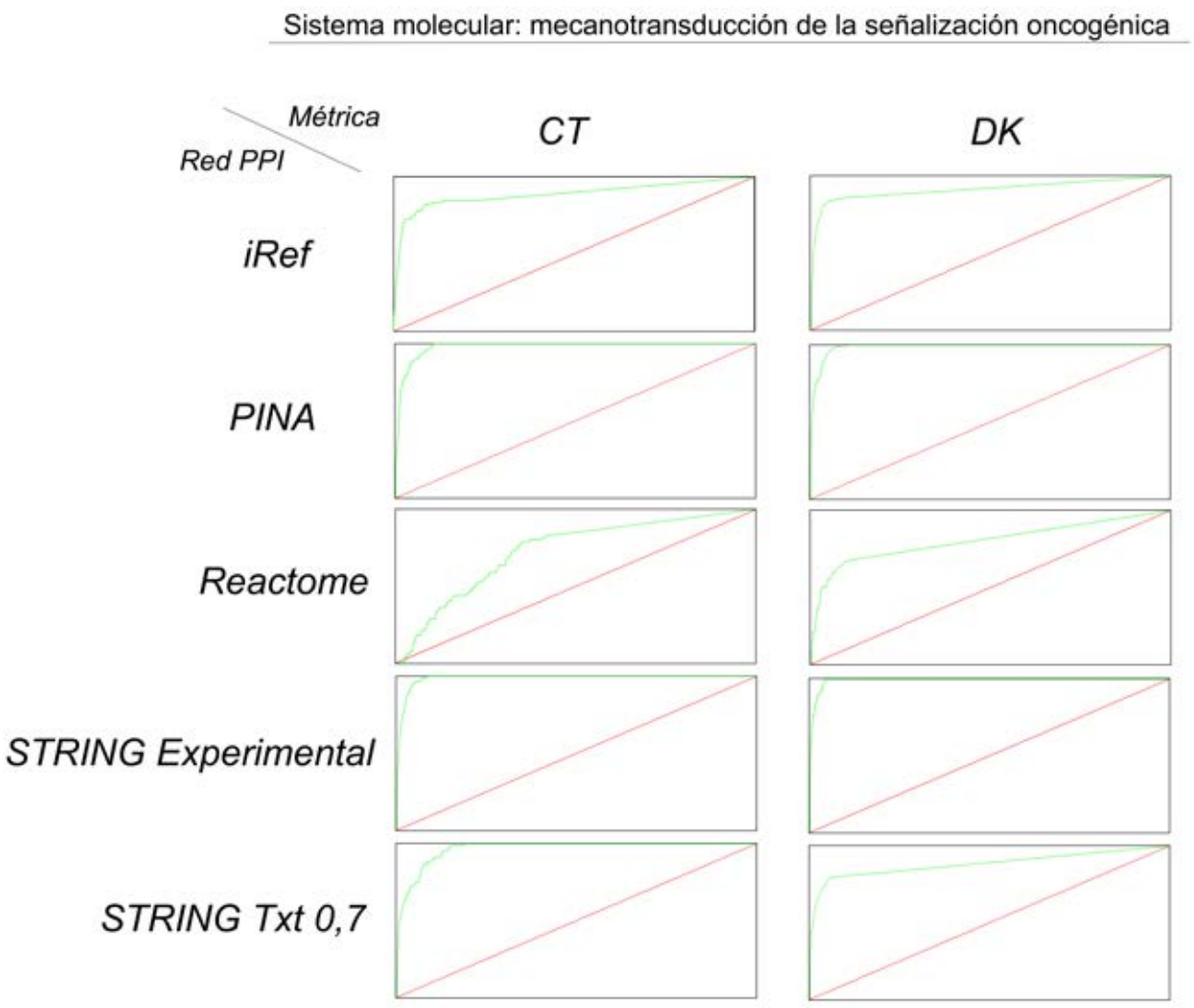


Figura 4.10: **Curvas ROC de los predictores para el sistema *mecanotransducción de la señalización oncogénica*.**
Misma descripción que en la figura 4.5.

Las figuras anteriores muestran los resultados de las pruebas realizadas a los predictores para todas las posibles combinaciones de red de interacciones y algoritmo de análisis de distancias en cada sistema molecular (*i.e.* adherencia célula-célula, regulación del citoesqueleto, etc.). Tal como se ha mencionado, esto se realizó con la finalidad de dilucidar qué combinación métrica/red tenía mayor capacidad predictiva para cada uno de los sistemas. Asimismo, se calcularon los valores del AUC para disponer de la comparativa de una forma numérica (véase la tabla 4.2).

Sistema molecular		**1**	**2**	**3**	**4**	**5**	**6**
Métrica	**Red PPI**						
CT	iRef	**0,909**	0,902	**0,881**	0,839	**0,857**	0,883
	PINA	0,814	0,895	0,811	0,985	0,759	0,971
	Reactome	0,581	0,781	0,596	0,897	0,657	0,722
	STRING Experimental	0,805	0,903	0,835	0,996	0,781	0,977
	STRING Textmining 0,7	0,729	0,729	0,640	0,865	0,676	0,957
DK	iRef	0,816	0,839	0,842	0,858	0,716	0,905
	PINA	0,819	0,915	0,832	0,987	0,767	0,976
	Reactome	0,715	0,787	0,612	0,936	0,689	0,796
	STRING Experimental	0,800	**0,926**	0,871	**0,998**	0,737	**0,981**
	STRING Textmining 0,7	0,663	0,700	0,595	0,670	0,717	0,875

Tabla 4.2: **Valores de AUC para cada uno de los predictores evaluados (red PPI y métrica), en cada sistema molecular.** La primera fila se corresponde con los 6 sistemas moleculares, en el siguiente orden: **1)** adherencia célula-célula, **2)** regulación del citoesqueleto, **3)** adhesión focal, **4)** ruta de señalización Hippo, **5)** regulación mecánica del núcleo y **6)** mecanotransducción de la señalización oncogénica. La primera columna agrupa los datos por métrica y la segunda por red de interacción de proteínas utilizada para la construcción del predictor. En negrita se muestran los valores máximos en cada sistema.

Tras este minucioso análisis se observó que existían 2 combinaciones de red de interacciones y algoritmo de medida que ofrecían los mejores rendimientos: la combinación de CT con la base de datos de interacciones iRef [21] y el algoritmo DK junto con STRING Experimental [23]. 3 de los sistemas tenían un mejor rendimiento con la primera combinación y otros 3 con la segunda (véanse los valores en negrita en la tabla 4.2). Los predictores que mostraron mejor rendimiento en cada caso (en cada sistema molecular), junto con sus valores de AUC, se pueden consultar en la tabla 4.3 y sus curvas ROC en la figura 4.11. Las puntuaciones de AUC obtenidas están en el rango de *buenas* a *excelentes* (véase el capítulo 3.3.2 de Materiales y métodos generales para una orientación sobre la interpretación de estos resultados). Dichos predictores fueron los seleccionados para obtener el resultado final, consistente en las listas ordenadas de genes/proteínas candidatos a formar parte de los sistemas moleculares, con sus valores de asociación funcional a los *sets de referencia*.

Sistema molecular	**Predictor seleccionado**	**AUC**
Adherencia célula-célula	CT - iRef	0,909
Regulación del citoesqueleto	DK - STRING Experimental	0,926
Adhesión focal	CT - iRef	0,881
Ruta de señalización Hippo	DK - STRING Experimental	0,998
Regulación mecánica del núcleo	CT - iRef	0,857
Mecanotransducción de la señalización oncogénica	DK - STRING Experimental	0,981

Tabla 4.3: **Combinaciones de red PPI y métrica seleccionadas para cada sistema molecular, por haber presentado el mejor rendimiento (mayor valor de AUC).** Para una interpretación de la calidad de los predictores en base al AUC se puede consultar el capítulo 3.3.2 de Materiales y métodos generales.

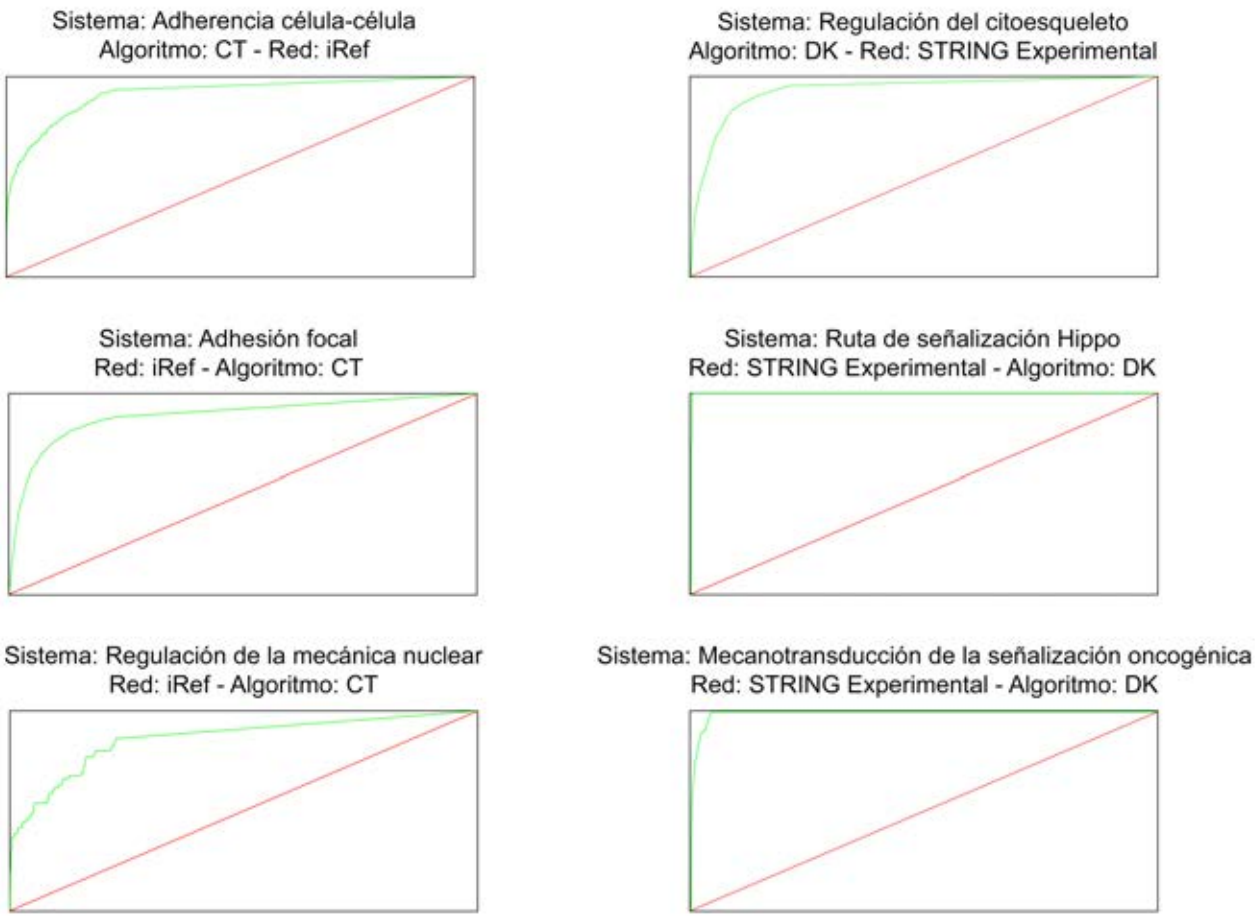


Figura 4.11: **Rendimientos de los predictores (en forma de curva ROC) seleccionados para cada uno de los sistemas moleculares por haber presentado los mejores resultados.** Tras analizar cada sistema usando cada una de las redes y cada uno de los algoritmos de medida, se seleccionó la mejor combinación de ellos (red y métrica) para ser usada como predictor. Los predictores seleccionados son los aquí mostrados para cada sistema. Un procedimiento basado en el método LOO fue usado para construir las curvas ROC y determinar el rendimiento. El eje *x* de la gráfica indica la tasa de falsos positivos y el eje *y* la tasa de verdaderos positivos. La diagonal, en rojo, representa el modelo aleatorio.

La aplicación de este procedimiento dio como resultado unos listados ordenados de proteínas candidatas según su afinidad funcional a los *sets de referencia* de los sistemas moleculares estudiados. De estos resultados es conveniente, posteriormente, hacer un cribado de las primeras posiciones de los listados mediante un análisis bibliográfico, guiado por expertos, con la finalidad de tratar de identificar si alguna de las proteínas seleccionadas mediante esta metodología ha sido ya estudiada en el contexto de estos sistemas moleculares; del mismo modo se aconseja un análisis bioinformático a través de herramientas como DAVID [18] o Gene Ontology (GO) [20], de manera que se puedan encontrar anotaciones funcionales o relaciones funcionales existentes. Tras ello, para alguna de las proteínas que han obtenido buenas puntuaciones con esta metodología y para las que no existe evidencia experimental de su vinculación a los sistemas moleculares asociados, se puede planificar un protocolo de experimentación *in vitro* con la finalidad de confirmar o desmentir la predicción realizada.

4.4.2. Discusión

En este trabajo se llevó a cabo un estudio predictivo con la finalidad de descubrir nuevas proteínas implicadas en procesos moleculares concretos. Más en detalle, se desarrolló la metodología necesaria para estudiar el caso de la transformación maligna de células tumorales de la línea MCF10CA1a cuando ocurrían cambios de rigidez en la matriz extracelular. Para ello se establecieron unas proteínas de referencia para cada proceso molecular implicado en este mecanismo celular, se seleccionaron diferentes representaciones del interactoma y se analizaron dichas redes de interacción mediante técnicas matemáticas, teniendo siempre como referencia las proteínas que formaban parte de dichos sistemas moleculares.

La validación teórico-matemática del método de predicción descrito en este capítulo se llevó a cabo mediante la técnica LOO y el cálculo de las curvas ROC y AUC. Los resultados de dichos tests mostraron unos rendimientos que oscilan entre buenos y excelentes. Por lo tanto, a nivel estadístico,

se puede afirmar que la metodología tiene un buen rendimiento y es robusta para los sistemas aquí estudiados.

La imposibilidad (debido a los problemas mencionados en la sección de antecedentes de este capítulo -véase 4.1-) de llevar a cabo una validación experimental de la implicación funcional de alguna de las proteínas predichas en los sistemas moleculares asociados al cambio a un fenotipo maligno de las células tumorales; hizo que se tomase la decisión de adaptar la metodología a otro caso de estudio donde sí se pudiese realizar la validación experimental de las predicciones.

Una de las propiedades de la metodología aquí descrita es su carácter transversal, ya que es aplicable a cualquier sistema o proceso molecular, representado en el interactoma humano, en el que se quiera realizar una predicción funcional sobre las proteínas que puedan estar implicadas en el mismo. Por lo tanto, se hizo una adaptación del método para la predicción de nuevas proteínas implicadas en la angiogénesis.

La angiogénesis (mecanismo de formación de vasos sanguíneos) se encuentra en el punto de mira de muchas investigaciones debido a que está relacionada con numerosas patologías [147]. La búsqueda de nuevos fármacos anti-angiogénicos es una estrategia terapéutica importante en cáncer, ya que en esta enfermedad la angiogénesis es uno de los factores cruciales para la progresión tumoral y la metástasis [148, 149]. Actualmente los fármacos conocidos en este sentido se centran en la inhibición de miembros de la familia de los factores de crecimiento endotelial vascular (VEGFs) y sus receptores, lo cual solo ha dado frutos en un limitado número de tipos de cáncer (como en los casos del cáncer colorrectal o del cáncer de mama), debido en parte a los mecanismos de resistencia observados en la mayoría de los tumores [14]. Es por ello que la identificación de nuevas dianas anti-angiogénesis incrementaría la posibilidad de diseñar nuevas terapias efectivas y con mayor espectro de aplicación.

La metodología bioinformática descrita en este capítulo fue aplicada a este nuevo caso de estudio, con algunas variaciones respecto a lo detallado en Material y métodos:

- El *set de referencia* fue constituido por 116 proteínas implicadas en angiogénesis, según el conocimiento experto y la bibliografía disponible.
- Las redes de interacción de proteínas fueron integradas en un solo modelo, mediante la fusión, principalmente, de las siguientes fuentes: CODA (predicciones funcionales a partir de datos de coincidencias de dominios protéicos) [25], GECO (similitud funcional derivada de datos de expresión génica) [26] y HIPPIE (relaciones de homología de interacciones) [27].
- El algoritmo de análisis de redes utilizado fue el RWR (*Random Walk with Restart*) por mostrar buen rendimiento en este caso. Véase el capítulo 3.2.1 de Materiales y métodos generales para más información.

Tras la construcción de los predictores y su validación mediante LOO y ROC se obtuvo un listado ordenado por significancia estadística compuesto por 19.618 proteínas candidatas, de las cuales se analizaron las 300 primeras -mediante un análisis bibliográfico por parte de expertos- y se decidió seleccionar 7 para llevar a cabo experimentos *in vitro*. Esta decisión se tomó en base a la novedad de la proteína candidata en la función angiogénica (se buscaron nuevas dianas no caracterizadas previamente) y a la viabilidad del diseño experimental y los costes asociados al mismo.

Se llevó a cabo la experimentación *in vitro* y *ex vivo*, haciendo uso de técnicas de silenciamiento génico y de bloqueo mediante anticuerpos específicos. Los resultados mostraron que la proteína SOD3 (*extracellular superoxide dismutase 3*) efectivamente estaba implicada en el proceso de angiogénesis. Más concretamente, su bloqueo reducía significativamente la migración de células endoteliales e inhibía completamente la formación de estructuras tubulares endoteliales; impidiendo la angiogénesis. Otras proteínas analizadas mostraron también indicios de estar implicadas en el proceso de angiogénesis pero los resultados no fueron tan concluyentes como en el caso de SOD3. Se realizaron experimentos adicionales *in vivo*, los cuales reforzaban estos hechos, apuntando a que SOD3 debería ser considerada como una nueva diana terapéutica en las patologías dependientes de la angiogénesis, tales como el cáncer.

Es por todo lo anterior que se puede afirmar que la metodología descrita en este capítulo fue validada a nivel teórico: mediante las curvas ROC y los valores del AUC mostrados, y que las predicciones obtenidas también se validaron a nivel práctico: a través de la investigación experimental llevada a cabo en el contexto del proceso de angiogénesis. La publicación asociada a dicha investigación se muestra en las siguientes páginas (capítulo 4.4.3).

Por último, se debe destacar el carácter extrapolable de esta metodología, la cual puede ser aplicada a cualquier proceso molecular de estudio. Asimismo, remarcar la utilidad de herramientas como la aquí detallada a la hora de aumentar la eficiencia de los diseños experimentales al focalizarlos sobre un grupo priorizado de proteínas enriquecidas en la función en estudio. Los silenciamientos génicos a nivel global (de todo el genoma) son, normalmente, demasiado caros para la mayoría de los grupos de investigación (al igual que otras técnicas de experimentación masiva), problema que se palía si llevamos a cabo los experimentos con un grupo más reducido de proteínas candidatas, descartando a la mayor parte de proteínas del interactoma completo.

4.4.3. Publicación: *In silico prediction of targets for anti-angiogenesis and their in vitro evaluation confirm the involvement of SOD3 in angiogenesis. Oncotarget 2018.*

La misma metodología descrita en este capítulo se adaptó para predecir nuevas proteínas implicadas en angiogénesis. Dichas predicciones fueron evaluadas posteriormente mediante experimentos de inhibición a través de *siRNA* y del bloqueo del producto génico con anticuerpos específicos. Tras la comprobación experimental de varias de las proteínas predichas por el sistema se confirmó la implicación en angiogénesis de SOD3 (*superoxide dismutase 3*), tanto *in vitro* como *ex vivo* e *in vivo*. De esta manera se han ratificado experimentalmente, con resultados positivos, predicciones realizadas mediante el método *in silico* descrito en este capítulo.

Fruto de esta investigación es el artículo que se adjunta.

Capítulo 5

Exploración de la red de interacciones de la familia de proteínas quinasas RAS en humanos y análisis de su evolución funcional orientado a potenciales implicaciones terapéuticas en cáncer

El universo no fue hecho a medida del hombre. Tampoco le es hostil. Es indiferente.

Carl Sagan

5.1. Introducción

La familia de proteínas RAS[1] es un conjunto de pequeñas enzimas GTP-asas monoméricas que están involucradas en la transducción de señales celulares, actuando como interruptores moleculares regulando la activación de redes de señalización intracelular. Controlan la homeostasis celular en eucariotas superiores mediante cambios conformacionales, activándose por diferentes factores de intercambio de guanosín trifosfato (*Ras Guanine Exchange Factor proteins* - RasGef) e inactivándose por proteínas activadoras de GTPasas (*Ras GTPase Activating proteins* - RasGap). De manera que funcionan como un conmutador molecular binario (*switch* o interruptor), alternando el estado activo unido a GTP e inactivo unido a GDP (véase la figura 5.1). Las rutas de señalización reguladas por RAS son responsables de funciones tales como: la integridad del citoesqueleto, la proliferación celular, la diferenciación funcional, la adhesión celular, la apoptosis -o muerte celular programada-, la migración celular y la supervivencia [28].

En el genoma humano, la familia *RAS* incluye un alto número de genes (parálogos), sin embargo, a excepción de unos escasos modelos proteicos bien estudiados, las funciones precisas de los 35 parálogos humanos *RAS* y su relación en términos de: conservación de secuencia, expresión génica e interacciones proteína-proteína siguen siendo prácticamente desconocidas [29].

Las mutaciones en los genes *RAS* y en genes codificantes de proteínas implicadas en la regulación de RAS: Ras, RasGef, RasGap, RapGap, etc. pueden ocasionar la transmisión inapropiada de señales en el interior de la célula, incluso en ausencia de señales extracelulares; del mismo modo, una desregulación en la señalización controlada por RAS puede derivar en procesos oncogénicos o cáncer. Hasta un 30 % del total de tumores humanos presentan mutaciones oncogénicas en miembros de la familia prototípica de RAS, induciendo patogénesis mediante la sobreactivación de la ruta

[1] El vocablo 'Ras' es escrito a lo largo del manuscrito en 3 formas diferentes, siguiendo estándares internacionales, dependiendo de su significado concreto: se hace uso de 'RAS' cuando se referencia de manera genérica a la familia de proteínas; se utiliza 'Ras' cuando se quiere aludir a una (o un conjunto) de las proteínas; y nombramos como '*RAS*' a los genes que dan lugar a las citadas proteínas.

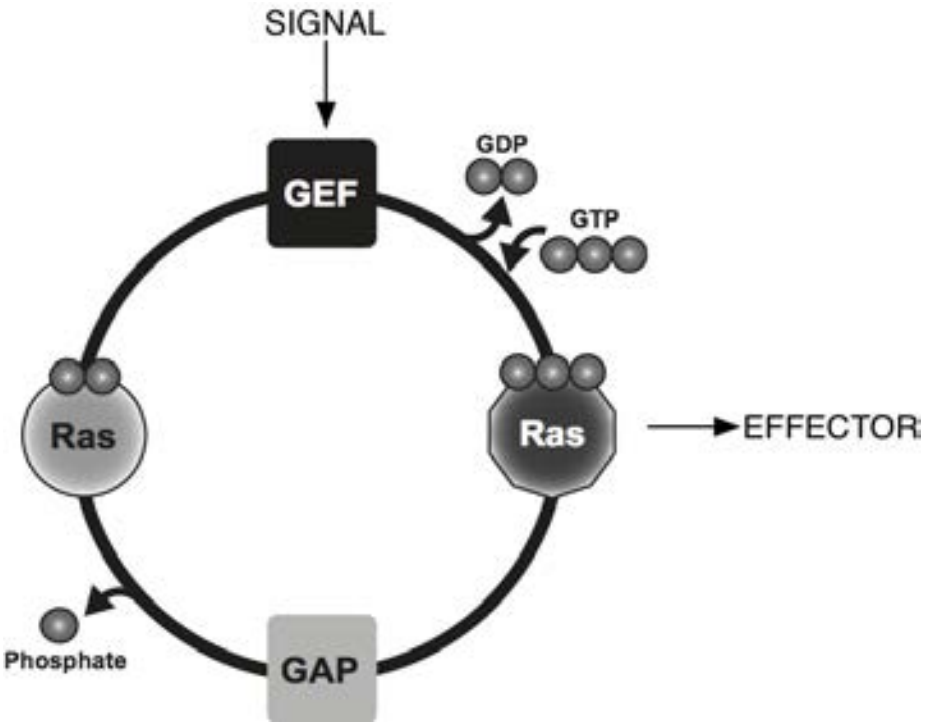


Figura 5.1: **Esquema de funcionamiento del *switch* (interruptor) de proteínas RAS.** La proteína Ras alterna 2 conformaciones estructurales: la forma activada donde se une al guanosín trifosfato (GTP) y la forma inactivada donde se une al guanosín difosfato (GDP). Imagen extraída de Díez *et al.* [29].

Raf/MEK/ERK [30, 31]. *KRAS* es el gen *RAS* que más frecuentemente se encuentra mutado (en más del 20 % de los tumores). Esto contrasta notablemente con el caso de los genes *NRAS* y *HRAS*, los cuales solo aparecen mutados en el 5 % y el 3 % de los casos, respectivamente. En particular, las mutaciones en *KRAS* predominan en los tumores pancreáticos, con una incidencia del 90 % (datos obtenidos del *Catalog Of Somatic Mutations In Cancer*, COSMIC, http://cancer.sanger.ac.uk/cosmic [150]). Las mutaciones oncogénicas de RAS se encuentran mayoritariamente en los residuos G12, G13 y Q61, obstaculizando la hidrólisis intrínseca de GTP y, por consiguiente, manteniendo a las proteínas RAS unidas a GTP, en su estado activo [151]. Además de al cáncer, mutaciones en los genes *HRAS* y *KRAS* han sido asociadas a los síndromes de Costello y Noonan, respectivamente [152, 153]. Cabe destacar que el resto de miembros de esta familia de genes no se encuentran significativamente mutados en cáncer y únicamente en algunos casos la sobreexpresión de genes

relacionados con *RAS* ha sido asociada a determinados tipos de tumores (i.e. RALA y RALB se sobreexpresan en melanoma y carcinoma pulmonar no microcítico (NSCLC), teniendo RALA un rol predominante en el crecimiento tumoral y RALB en su potencial metastático) [154, 155].

Tal como se ilustra en la figura 5.2, la red humana de proteínas RAS puede cambiar su topología a través de 2 mecanismos básicos: i) cambiando los nodos presentes en la red (i.e. mediante cambios en la expresión génica); y ii) reconfigurando las conexiones entre los nodos (i.e. mediante mutaciones en las interfaces de unión entre las proteínas).

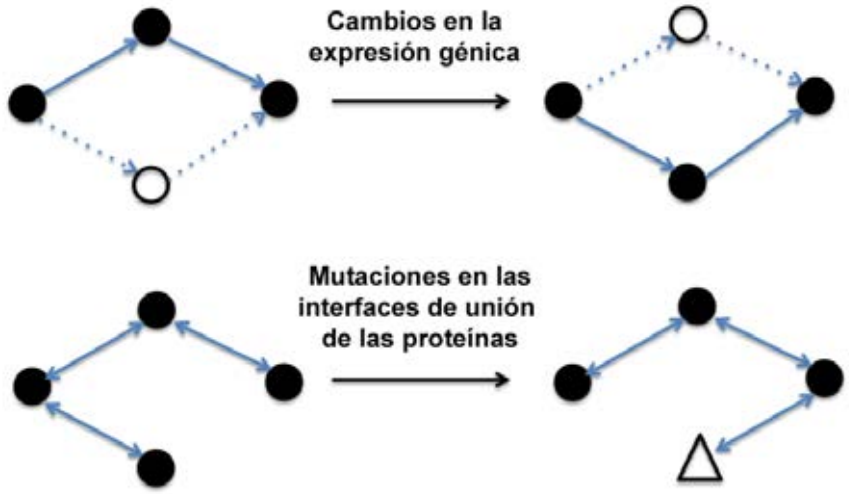


Figura 5.2: **Ejemplo de diferentes mecanismos biológicos que cambian la topología de la red de interacciones.** En la parte superior, donde los nodos negros representan proteínas que se expresan y las líneas continuas interacciones activas, se muestra el efecto de los cambios en la expresión génica. Por otro lado, la parte inferior representa el efecto de un proceso de reconfiguración inducido por mutaciones (triángulo) en los interfaces de unión de las proteínas.

Así como las 3 proteínas RAS prototípicas han sido ampliamente estudiadas y extensamente caracterizadas, mucho menos se conoce acerca de los restantes parálogos RAS, tanto en tejidos sanos como tumorales. En este trabajo se estudia la relación entre las distancias filogenéticas de los parálogos RAS y sus asociaciones en la red de interacciones del proteoma humano (véase la figura 5.3 donde se detalla este proceso). Adicionalmente, se ha implementado un análisis comparativo

de secuencias con el fin de encontrar posiciones de aminoácidos significativamente conservadas entre pares de proteínas RAS divergentes filogenéticamente que conserven la proximidad en el interactoma humano. La hipótesis es que estas posiciones pueden haber ayudado a mantener las interacciones entre proteínas funcionalmente importantes y comunes para ambos parálogos, resultando en su proximidad en la red de interacciones. Estas posiciones son, posteriormente, mapeadas en diferentes complejos de proteínas RAS con otras proteínas usando la información de sus estructuras tridimensionales con el fin de determinar su papel en los sitios de unión de las proteínas RAS con las otras proteínas con las que interaccionan.

Los resultados que aquí se muestran añaden una nueva perspectiva a la idea generalmente aceptada de que las interacciones entre las proteínas parálogas divergen con su secuencia [32–34] y arrojan algo de luz sobre el papel ampliamente desconocido de la red de interacciones de RAS en humanos. Además, estos hallazgos amplían la visión actual sobre el papel putativo de los genes parálogos en el desarrollo y la adaptación de redes de señalización RAS funcionales y patológicas. Adicionalmente, se pueden sacar conclusiones importantes sobre las posiciones conservadas en los *Divergent but Interacting RAS Pairs* (DIRP). DIRP hace referencia a aquellos pares de proteínas distantes filogenéticamente pero cercanos en la red de interacciones. En este trabajo se desarrolla un análisis en profundidad de la posible relevancia funcional de las posiciones conservadas en los DIRP en la mediación de la interacción de RAS con sus efectores, lo cual tiene implicaciones en el diseño y desarrollo de nuevos inhibidores de Ras con fines terapéuticos.

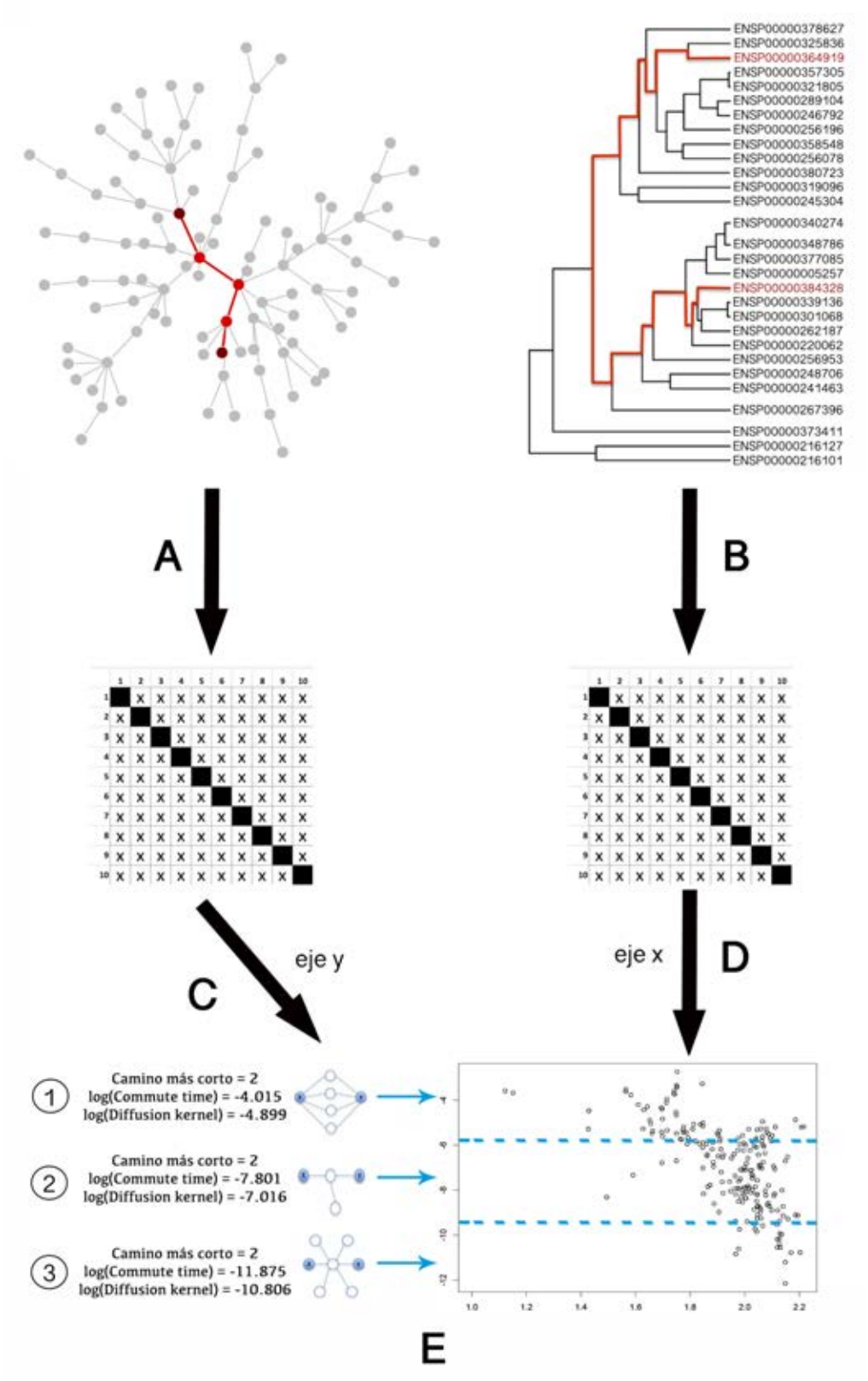


Figura 5.3: **Esquema del proceso de comparación de medidas de distancia en red y filogenéticas entre pares de proteínas Ras.** Panel **A)** Cálculo de las distancias entre pares dentro del grafo de PPI, expresadas como una matriz. Panel **B)** Cálculo de las distancias filogenéticas entre pares en el árbol, expresadas como una matriz. Panel **C)** Transformación logarítmica para normalizar las distancias en red entre las proteínas. Panel **D)** Transformación exponencial para normalizar las distancias filogenéticas entre las proteínas. Panel **E)** Representación gráfica conjunta de las distancias filogenéticas y en red. Tal como se observa en la parte izquierda del panel E, las medidas de distancia basadas en *kernels* (e.g. DK o CT), al contrario que la del cálculo de la distancia más corta (mínimo número de aristas conectando 2 nodos), son capaces de distinguir el nivel de asociación entre 2 nodos conectados mediante diferentes topologías: 1) nodos altamente conectados; 2) escasamente conectados; 3) conectados de manera no específica. Este resultado demuestra que las medidas de similitud basadas en técnicas *kernel* son una de las mejores herramientas a la hora de identificar diferentes contextos de interacciones.

5.2. Material y métodos

Para responder a las cuestiones planteadas con respecto a la familia de proteínas RAS, y más concretamente a la relación entre su funcionalidad y forma de evolución, se han extraído por un lado los datos de distancias filogenéticas (véase el capítulo 3.1) y por otro se han estudiado, mediante diferentes métodos de análisis, las redes de interacción de proteínas (véase la figura 5.4 y el capítulo 3.2.1).

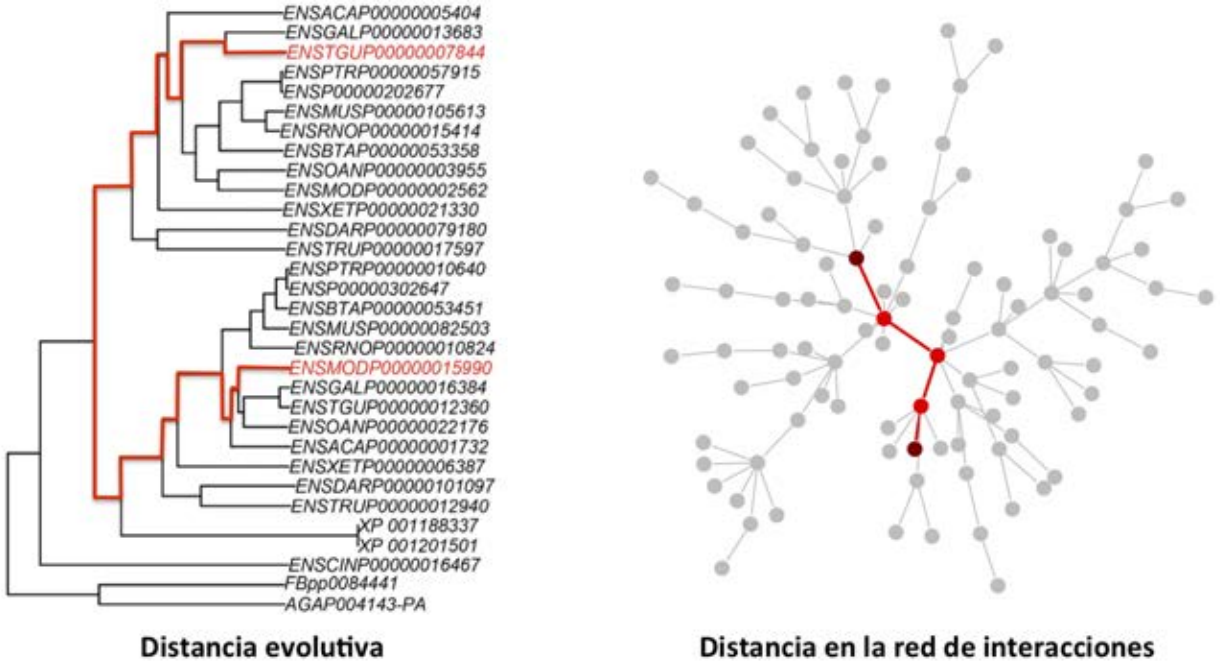


Figura 5.4: **Distancia entre 2 proteínas en un árbol filogenético y en una red de interacción de proteínas.**

5.2.1. Árboles filogenéticos de la familia Ras

Los árboles filogenéticos de las 35 proteínas parálogas humanas de Ras usadas en este trabajo fueron parte del conjunto de datos obtenido en Diez *et al.* [29]. Estos árboles fueron el producto de

una precisa y exhaustiva búsqueda de todos los genes codificantes de proteínas de la familia Ras a lo largo de 24 especies eucariotas.

Diez *et al.* [29] obtuvieron las secuencias de Ras de Uniprot [156] y las alinearon en un MSA (*Multiple Sequence Alignment* o alineamiento múltiple de secuencias) con sus ortólogos haciendo uso de ClustalW [35]. Finalmente, los árboles filogenéticos fueron construidos mediante el método *Neighbor-Joining* (véase la figura 5.5), haciendo uso del software Quicktree [36]. La fiabilidad de la topología del árbol se evaluó con el método *bootstrap* usando 1000 replicaciones.

Figura 5.5: Proceso de construcción de los árboles filogenéticos utilizados.

5.2.2. Datos de las redes de interacción proteína-proteína

Para la selección de las redes de interacción de proteínas a utilizar, de entre todas las detalladas en el capítulo 3.2, se tuvieron en cuenta los siguientes puntos:

1. Que los datos de las redes no proviniesen de fuentes filogenéticas, para evitar resultados tautológicos al compararlas con los árboles.
2. Que el nivel de cobertura (proteínas presentes tanto en los árboles filogenéticos como en las redes de interacción) fuese el mayor posible.

Las 2 redes de interacción de proteínas analizadas finalmente en este trabajo fueron: STRING [23] y PINA [22]; concretamente las siguientes versiones: STRING - canal experimental v8.3 y

PINA v.20110225. El uso de 2 redes se justifica en la búsqueda de una mayor robustez en los resultados.

STRING [23] describe, en el momento de llevar a cabo este trabajo, 263.666 interacciones entre 14.732 proteínas. En este estudio se hizo uso únicamente de interacciones físicas directas (*escore* -véase el capítulo 3.2 de Materiales y Métodos generales para más información acerca de la estructura de STRING y el filtrado por canales según la procedencia de los datos-), evitando tanto los datos derivados de los estudios filogenéticos (para prevenir tautologías en los resultados cuando son comparados con las distancias en los árboles) como los obtenidos mediante procesos de minería de textos científicos [157], los cuales no distinguen entre relación funcional o interacción física directa.

PINA [22] incluye 108.477 interacciones únicas entre 15.450 proteínas diferentes y cubre el 63 % de las proteínas presentes en el árbol filogenético de Ras y el 31 % de todas las posibles conexiones entre ellas, mientras que STRING cubre el 77 % y 52 % respectivamente. Aunque los datos de PPI de PINA y STRING se obtienen integrando fuentes de información similares (interacciones físicas, tal como se ha mencionado), muestran un nivel de cobertura diferente con los datos del árbol de Ras y también una topología de red distinta. Por lo tanto, ambos fueron considerados como conjuntos de datos válidos y complementarios en este análisis.

5.2.3. Distancias entre pares en las redes PPI y los árboles filogenéticos

Las proteínas RAS fueron mapeadas en las redes PPI y los nodos altamente conectados (aquellos con 300 o más conexiones) fueron eliminados, debido a que estos nodos promiscuos introducen ruido en los cálculos de distancias, tal como muestra Hériché *et al.* [46]. De entre todos los algoritmos probados, el *Laplacian Exponential Diffusion Kernel* (DK) y el *Commute Time Kernel* (CT), fueron los que mejor se adaptaron a los propósitos de este estudio (véase el capítulo 3.2.1). Por lo tanto, las distancias entre pares de proteínas dentro de las redes fueron calculadas haciendo uso de

estos métodos. Ambos están basados en el cálculo de la probabilidad (p) de asociación de pares de nodos en la red utilizando diferentes aproximaciones estadísticas para representar matemáticamente el flujo en la red (véase el capítulo 3.2.1). Cabe destacar que CT también está incluido como parte de herramientas ampliamente utilizadas, tales como GeneMANIA [158]. Estas probabilidades (p) fueron normalizadas y transformadas en distancias calculando su logaritmo natural negativo ($-Ln(p)$), para contrarrestar su distribución exponencial (véase el panel A en la figura 5.6)[2].

Por otro lado, las distancias filogenéticas entre pares fueron calculadas haciendo uso del algoritmo descrito por Pazos *et al.* [45], el cual utiliza los archivos de los árboles de proteínas en formato *Newick Standard* [90] (véase el capítulo 1.2) como entrada y devuelve el valor numérico de distancia para cada par. Posteriormente se llevaron a cabo correcciones de escala, aplicando una transformación matemática exponencial a las distancias filogenéticas, de manera que pudiesen ser representadas y comparadas de forma conjunta con las distancias en la red (en el panel B de la figura 5.6 se muestra la distribución logarítmica de los valores en el árbol previa a la normalización).

La figura 5.1 evidencia la necesidad de ambas normalizaciones antes de la comparativa.

[2] A lo largo de este capítulo se presenta una serie de figuras con fondo gris bajo el título 'Material suplementario'. Se trata de figuras que originalmente forman parte del material adicional (sección 5.3.7) de la publicación fruto del trabajo detallado en este capítulo, pero que se ha estimado conveniente incluir en la parte principal del capítulo.

Material suplementario

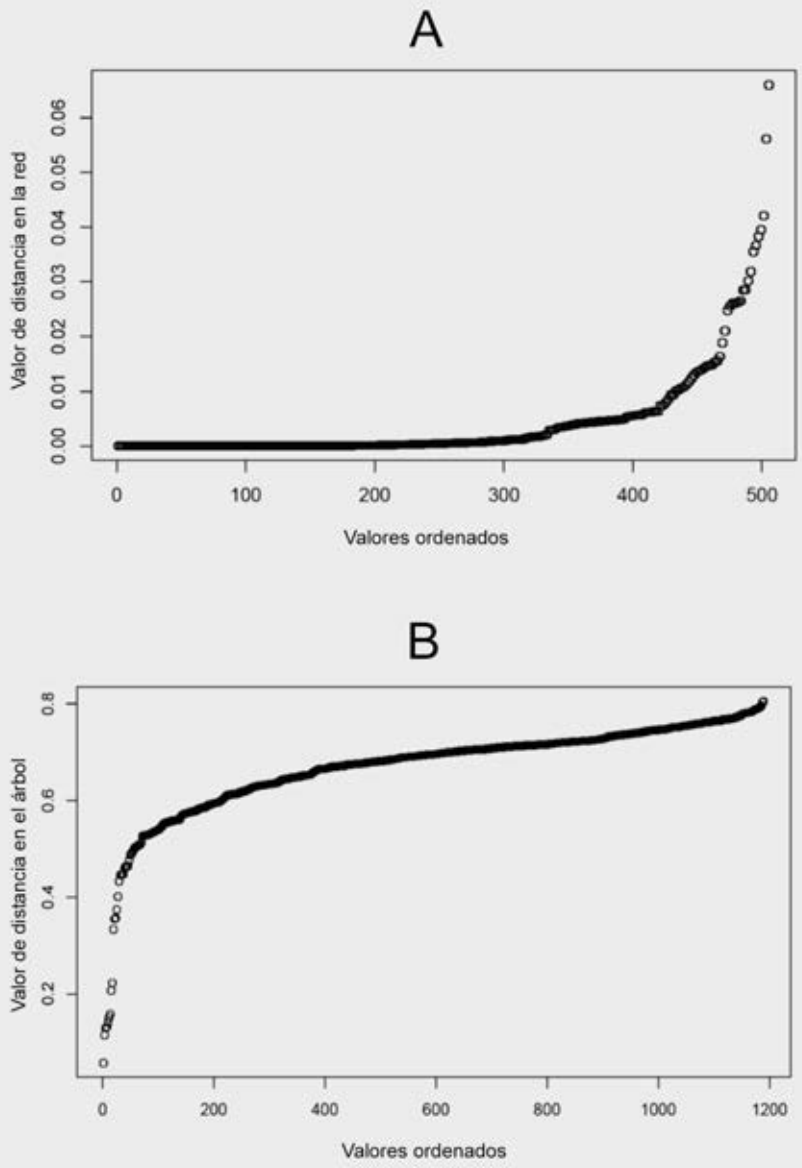


Figura 5.6: **Ejemplo de distribución de los valores de distancia en red y de similitud filogenética.** A) Distancias entre pares (CT) de proteínas Ras en *Homo sapiens* en la red STRING Experimental. B) Distribución de distancias filogenéticas entre pares del mismo conjunto de proteínas. La naturaleza intrínseca de los datos obtenidos de CT y DK (transformaciones no lineales) los distribuye de manera exponencial, mientras que en los derivados de las distancias filogenéticas se observa un comportamiento semejante al logarítmico.

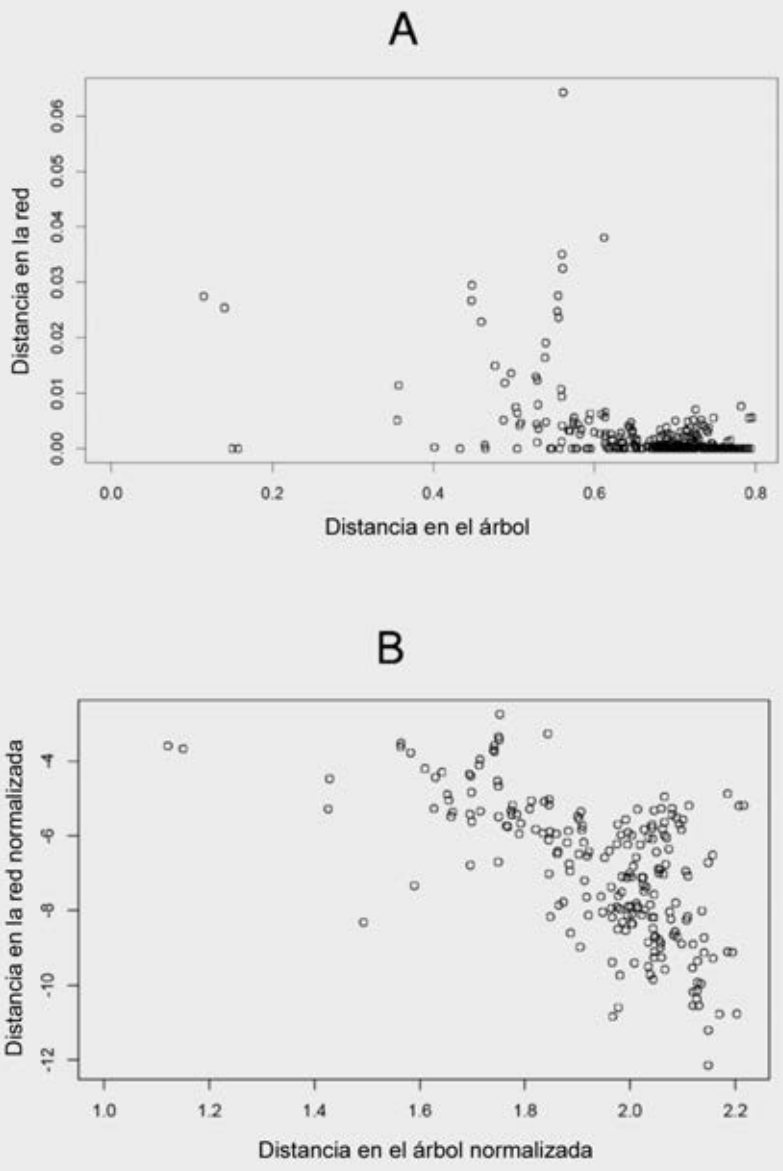


Figura 5.7: **Efecto de la normalización de medidas en la comparativa entre distancias en la red y distancias filogenéticas.** Distancia filogenética entre pares (eje x) *vs.* distancia en la red entre pares (eje y) antes (A) y después (B) de la normalización.

Finalmente, la comparación entre las distancias filogenéticas y las matrices de las redes PPI, así como las representaciones gráficas correspondientes fueron llevadas a cabo haciendo uso del software R [47].

5.2.4. Selección de los pares RAS divergentes pero interactuantes (DIRP)

Para seleccionar los pares de secuencias divergentes, se definió un punto de corte de identidad máxima del 45 %. Este valor se basó en la matriz BLOSUM 45 [48], la cual fue diseñada como estándar para medir sustituciones de aminoácidos entre secuencias altamente divergentes. Este umbral seleccionado se corresponde a una distancia filogenética normalizada entre proteínas mayor a 1,7.

Con la finalidad de establecer una cercanía significativa entre proteínas en las redes de interacción, se fijó un segundo punto de corte basado en distribuciones aleatorias de los valores de distancia en red de DK y CT. Para cada conjunto de datos (red) y algoritmo, este punto de corte fue estimado estadísticamente de acuerdo a un p-valor = 0,05 (el valor numérico concreto se muestra en la tabla 5.1).

Material suplementario

PPI network dataset / Normalized network distance boundaries	$CT_{0.05}$	$DK_{0.05}$
PINA	-6.5	-12
STRING EXP	-6.5	-11.5

Tabla 5.1: **Puntos de corte para la selección de la distancia -cercanía- en red significativa.** Las columnas $CT_{0,05}$ y $DK_{0,05}$ representan los valores de punto de corte en las distancias en la red (normalizados logarítmicamente) para un p-valor de 0,05.

Finalmente, aquellos pares con una identidad de secuencia $\leq$ 45 % y valores de DK y CT $\geq$ $DK_{0,05}$ y $CT_{0,05}$, respectivamente, fueron utilizados para construir el conjunto final de pares DIRP (*Distant but Interacting Ras Pairs*; pares alejados filogenéticamente pero cercanos en la red de interacciones).

5.2.5. Alineamiento múltiple de secuencias y medida de la conservación de los aminoácidos

Un alineamiento múltiple de todas las secuencias Ras (MSA) fue empleado para medir la conservación de aminoácidos en cada posición de las proteínas. Esta evaluación fue llevada a cabo usando la matriz de sustitución de aminoácidos BLOSUM 45 para puntuar todos los cambios en cada posición de las secuencias. La elección de BLOSUM 45 se basó en el hecho de que esta matriz fue originalmente diseñada para comparar secuencias altamente divergentes, con hasta un 45 % de identidad; una condición que este conjunto de datos cumple mayoritariamente. Únicamente aquellos aminoácidos que alineaban con la secuencia de la proteína HRas fueron usados para el análisis de conservación. HRas fue seleccionada como plantilla de referencia por ser la proteína más estudiada de la familia y una de las principales dianas farmacológicas.

Para cada posición (aminoácido) en el MSA se calcularon 2 valores: i) la media del nivel de conservación entre pares DIRP (control positivo) basado en alineamientos binarios entre todos los pares del conjunto de DIRP y ii) la media del nivel de conservación de un mismo número de pares de proteínas Ras seleccionadas aleatoriamente (control negativo) haciendo uso de la misma aproximación que en 'i)'. Estos 2 valores fueron entonces normalizados haciendo uso de la media de conservación en el MSA global. En base a los resultados de los modelos aleatorios (control negativo), se calculó el p-valor asociado y se usó como punto de corte para seleccionar los aminoácidos significativamente conservados entre los pares DIRP (p-valor $\leq$ 0,01). El flujo general del proceso se muestra en los puntos A, B y C de la figura 5.8.

Tanto la visualización como la edición del MSA se llevaron a cabo mediante el software *JalView* V2.7 [159].

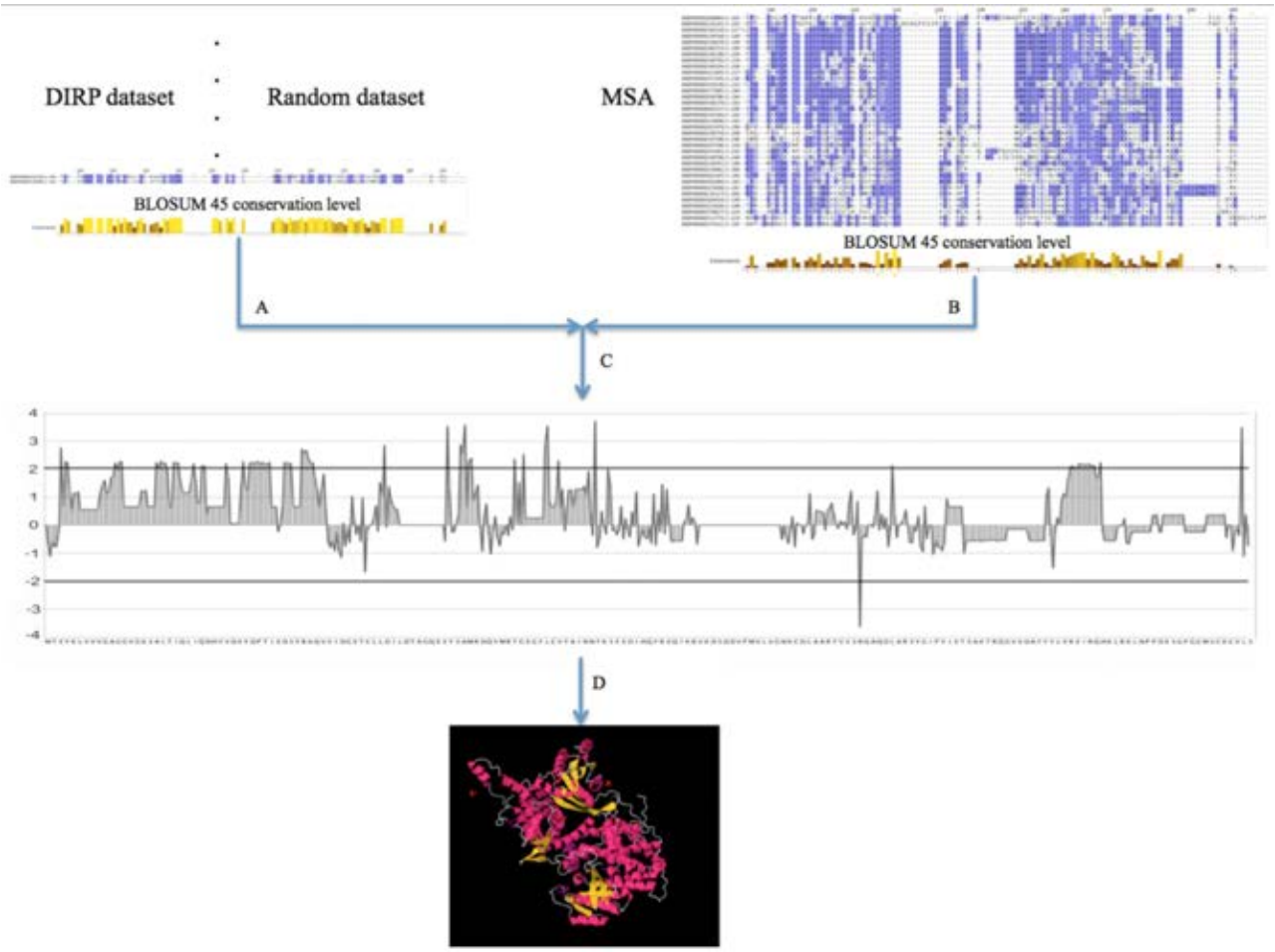


Figura 5.8: **Proceso general para la obtención del conjunto de posiciones específicas en los DIRP y su mapeo en los complejos 3D de Ras. A)** Medición de la conservación de la posición haciendo uso de la matriz BLOSUM 45 para los pares seleccionados como DIRP y para los pares seleccionados aleatoriamente (control negativo). **B)** Medida de la conservación de las posiciones en el MSA completo. **C)** Cálculo de la variación de la conservación para cada posición (normalización) entre el conjunto de pares DIRP y el aleatorio, comparada con la conservación de dicha posición en el MSA. **D)** Selección de las posiciones específicas con diferencia de conservación significativa en los DIRP (p-valor $\leq$ 0,01 en base al modelo aleatorio) y su mapeo en los diferentes complejos 3D de unión de Ras en *Homo sapiens*.

5.2.6. Modelos aleatorios

Para cada red PPI y algoritmo utilizado, se generaron modelos aleatorios en diferentes etapas del trabajo con la finalidad de estimar la significancia estadística de los resultados (*i.e.* para ser usados como controles negativos).

5.2.7. Modelos aleatorios del interactoma

Se construyeron cien modelos aleatorios de red PPI para cada una de las redes PPI utilizadas, permutando aleatoriamente los vecinos de cada nodo mientras se mantenía su grado de conexión. Las distancias en la red fueron posteriormente calculadas en estos modelos y comparadas con las distancias filogenéticas.

5.2.8. Conjunto aleatorio de pares de proteínas alineadas

Se construyeron cien conjuntos de pares de secuencias de proteínas alineadas seleccionadas aleatoriamente del alineamiento múltiple de secuencias (MSA). Los tamaños (en número de pares) de los conjuntos aleatorios se mantuvieron idénticos a los del grupo de datos original (DIRP).

5.2.9. Adquisición y procesado de los datos estructurales de los complejos Ras

Todos los complejos 3D de interacción conocidos de las proteínas humanas Ras fueron descargados de *Protein Data Bank* (PDB) [49]. Aquellos con una identidad de secuencia del 100 % fueron agrupados conjuntamente (véase la tabla S3 del material suplementario -capítulo 5.3.7- y el enlace: goo.gl/wcdiKV) y posteriormente etiquetados en categorías funcionales según su similitud

en cuanto a estructura tridimensional (r.m.s.d. $<$ 1,0; véase la tabla S4 del material suplementario -capítulo 5.3.7- y el enlace: goo.gl/wcdiKV). Para cada grupo funcional, la superficie de interacción de Ras fue determinada mediante la computación de la diferencia de la superficie accesible al solvente de los aminoácidos de Ras entre el complejo y los estados disociados, haciendo uso del software DSSP [50]. Los datos relativos a las frecuencias de mutación fueron obtenidos de COSMIC (http://cancer.sanger.ac.uk/cosmic) [160]. Los modelos estructurales fueron representados visualmente haciendo uso de *The PyMOL Molecular Graphics System*, versión 1.8 Schrödinger, LLC.

5.3. Resultados y Discusión

5.3.1. Relaciones entre las distancias en red y las filogenéticas de los parálogos *RAS* en humanos

Para analizar la relación entre la filogenia de las proteínas RAS y su localización en las redes de interacción proteína-proteína (PPI), se compararon las distancias obtenidas en la red y en el árbol para todos los pares de parálogos RAS en humanos (véase la figura 5.3 y Material y métodos -capítulo 5.2-). Los parálogos RAS mostraron tendencia a estar cerca en el interactoma cuando eran filogenéticamente cercanos y a incrementar su distancia en la red de interacciones a medida que presentaban mayor divergencia en el árbol. Se observa el mismo patrón independientemente de la red PPI y la medida de distancia utilizada (véase la figura 5.9). Este patrón no se encontraba en los modelos aleatorios (véase la sección 5.2 -Material y métodos de este capítulo-). Tal como se observa en la figura 5.9, las distancias en la red de los pares más divergentes se asemejaban a una distribución aleatoria, mientras que los pares filogenéticamente cercanos mostraban una distribución de distancias en la red muy distinta a la esperada por azar.

La correlación inversa entre la similitud de secuencia y la distancia filogenética de los pares de proteínas Ras es consistente con un modelo evolutivo clásico por el cual los genes recientemente duplicados comparten el mismo contexto de interacciones. Por lo tanto, a medida que las secuencias divergen por acumulación de mutaciones, se alejan entre sí en el interactoma. Sin embargo, los resultados aquí presentados muestran además que algunos de los genes duplicados distantes filogenéticamente mantienen el mismo contexto en la red de interacciones proteína-proteína, sugiriendo que hay algo más afectando al modelo evolutivo clásico antes mencionado.

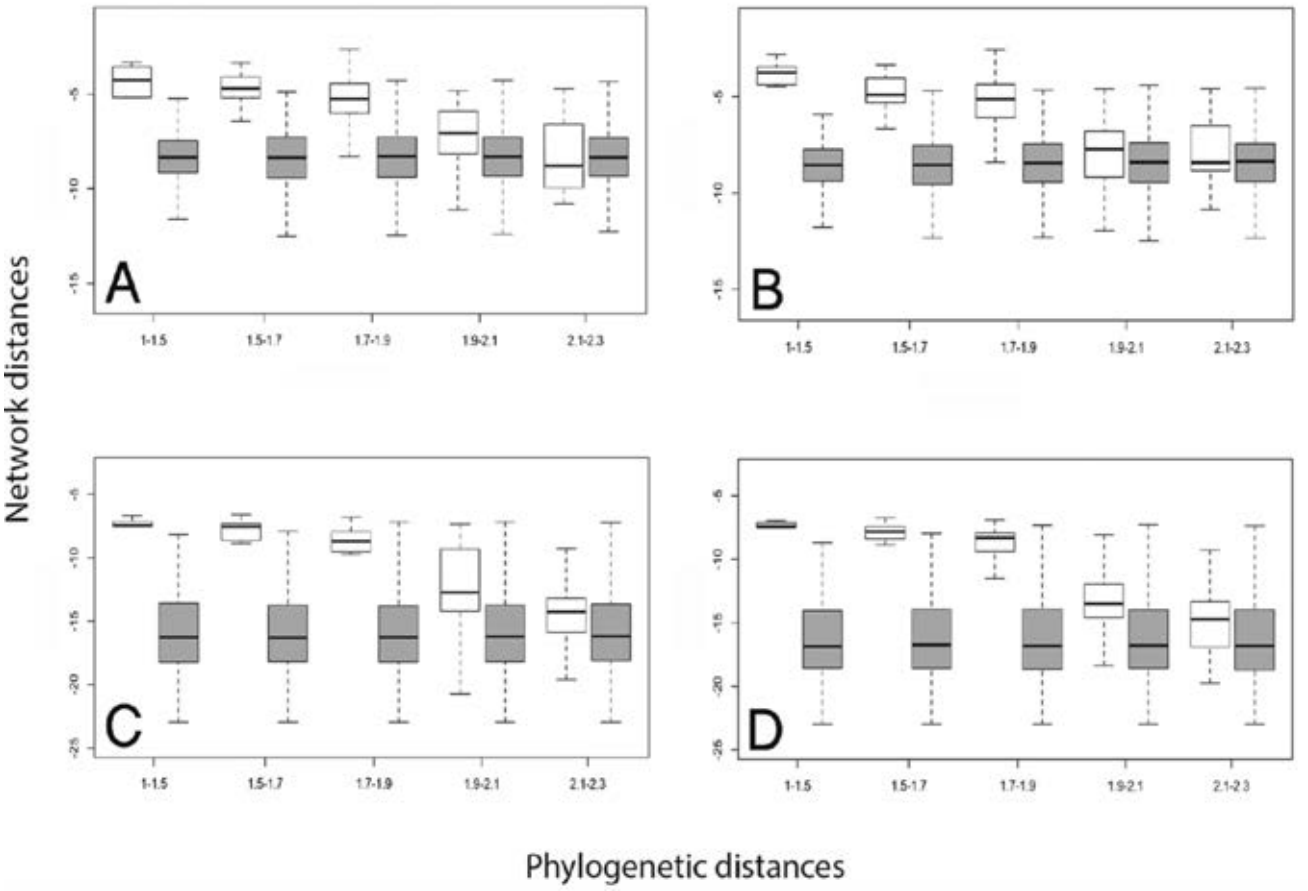


Figura 5.9: **Distribución de las distancias en la red de interacciones entre pares de proteínas en función de la distancia filogenética.** Distribuciones normalizadas de las distancias en red entre pares de proteínas Ras (eje y), divididas en segmentos que corresponden a rangos crecientes de distancias filogenéticas normalizadas (eje x). Las cajas blancas se corresponden con los valores reales y las oscuras con los valores en los modelos aleatorios. Las distancias en la red de interacciones fueron calculadas aplicando los algoritmos CT (paneles **A** y **B**) y DK (paneles **C** y **D**) en los grafos STRING Experimental (paneles **A** y **C**) y PINA (paneles **B** y **D**).

5.3.2. Identificación de pares de parálogos Ras divergentes con localización cercana en la red PPI (DIRPs)

Como se acaba de mencionar, existe una correlación inversa entre la conservación de secuencia y la distancia filogenética en los pares de proteínas Ras. De esto se puede también concluir una relación inversa entre la conservación de secuencia y la distancia en red en base a los resultados mostrados en la figura 5.9. Esta observación sugiere que la conservación o variación de aminoácidos en distintas posiciones de la secuencia podría determinar si un par de proteínas Ras comparte los mismos vecinos con los que interacciona o difiere en ellos, dentro de una red PPI. Con el objetivo de identificar estas posiciones aminoacídicas potencialmente determinantes de la localización de las proteínas Ras en la red PPI, se examinó detalladamente la relación entre las distribuciones de las distancias filogenéticas y las distancias en la red PPI de todos los pares Ras (figura 5.9). Se distinguen 4 áreas principales en los gráficos comparativos de distancias filogenéticas *vs.* distancias en la red, basados en 2 valores utilizados como umbrales, uno para las medidas de distancias en la red y otro para las distancias filogenéticas (véase tanto el capítulo 5.2.4 de Material y métodos como las áreas I-IV en la figura 5.10). Estas 4 zonas son: **área I)** Pares Ras cercanos en el árbol filogenético y en el grafo de la red PPI: en esta zona la alta conservación generalizada entre secuencias hace difícil distinguir aquellas posiciones conservadas responsables de la localización cercana en la red PPI que se observa en este conjunto de pares; **área III)** Pares Ras cercanos en el árbol y distantes en la red PPI: esta zona se encuentra vacía, sugiriendo que unas pocas mutaciones en genes *RAS* recientemente duplicados no puede producir un cambio sustancial en sus contextos de interacción con otras proteínas en la red; **área IV)** Pares Ras distantes en el árbol y en la red PPI: en esta zona IV la alta divergencia entre secuencias de nuevo hace difícil identificar aquellas posiciones variables directamente responsables de la divergencia en los contextos de interacción de estos pares; finalmente, el **área II)** Pares Ras distantes en el árbol pero cercanos en la red PPI: en esta zona encontramos un conjunto de pares con secuencias divergentes donde sería factible identificar posiciones conservadas específicas, relacionadas con su localización cercana en la red PPI. Se hace referencia a este conjunto de pares de parálogos como DIRP (*Divergent but Interacting RAS Pairs*).

Con la finalidad de diferenciar los datos significativos de aquellos comportamientos aleatorios sin relevancia, el conjunto de datos DIRP fue seleccionado de entre todos los pares RAS teniendo en cuenta 2 filtros estadísticos de significancia: i) que la divergencia entre las secuencias de las proteínas en el par fuese significativa y ii) que la cercanía en el interactoma de proteínas fuese significativa (véase Material y métodos -capítulo 5.2.4- y la figura 5.10).

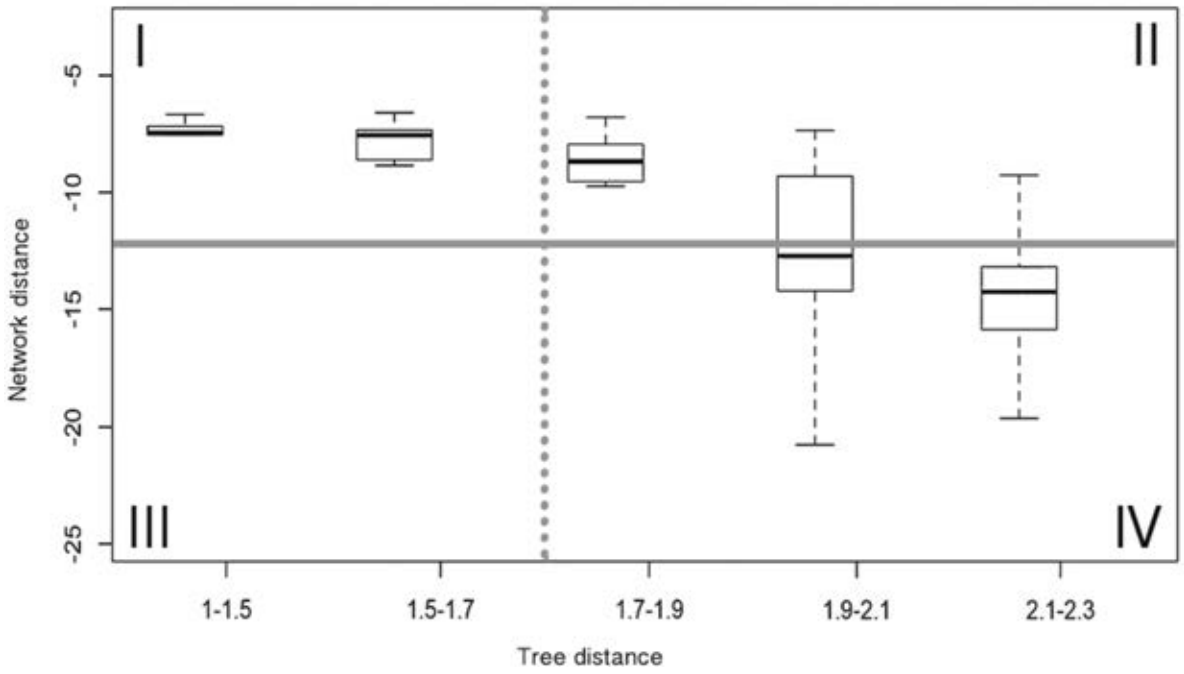


Figura 5.10: **Distribución de los valores de distancias en red *vs.* distancias filogenéticas y los puntos de corte establecidos.** Ejemplo de la comparación entre las distancias filogenéticas normalizadas y las distancias en red normalizadas entre pares de proteínas, aplicando el algoritmo DK al conjunto de datos STRING Experimental para la obtención de las distancias en la red. El punto de corte en las distancias filogenéticas corresponde a pares con un 45 % de identidad de secuencia (línea discontinua) y el punto de corte de cercanía en la red se establece de acuerdo a un p-valor de 0,05 (línea continua).

El número de pares de proteínas que finalmente fue incluido como DIRP se muestra en la tabla 5.2, para cada modelo de red PPI y cada métrica de distancia en red utilizada.

Se estudió la relación entre la cercanía en la red y la similitud de los contextos de interacción en el conjunto de datos DIRP, analizando todos los interactores (terceras proteínas) compartidos

	DK STRING Exp	CT STRING Exp	DK PINA	CT PINA
# initial pairs	351	351	435	435
#pairs after phylogenetic boundary	323	323	396	396
# DIRP	106	82	113	86
% pairs	30	23	26	20

Tabla 5.2: **Número de pares de proteínas Ras a lo largo de todo el proceso de selección para la obtención de los DIRP.** La primera fila indica el número de pares de proteínas que fueron inicialmente analizados en cada sistema (algoritmo y conjunto de datos utilizado). La segunda fila muestra el número de pares después de aplicar el corte de significancia basado en la distancia filogenética para pares lejanos (distancia filogenética normalizada $\geq$ 1,7). La tercera fila contiene el número de pares DIRP seleccionados finalmente, tras filtrar mediante el punto de corte en las distancias en red normalizadas (p-valor $\leq$ 0,05) establecido por medio de modelos aleatorios y especificado en la tabla 5.1 de Material y métodos. La última fila indica los porcentajes de pares DIRP sobre el número total de pares Ras obtenidos inicialmente.

directamente para cada uno de los pares Ras en este grupo DIRP, y comparándolos con un conjunto de datos equivalente No-DIRP (véase la figura 5.11). Los pares DIRP mostraron una mediana de 3 interactores (proteínas) compartidos por par, mientras que en el grupo No-DIRP la mediana era de cero. Los resultados son prácticamente los mismos haciendo uso de las métricas de similitud *Commute Time Diffusion Kernel* (CT) o *Laplacian Exponential Diffusion Kernel* (DK). Estos resultados prueban la existencia de una correlación positiva entre el número de proteínas interactuantes compartidas por los pares DIRP y la cercanía en la red medida con métricas *kernel*. Un detallado análisis de algunos de los pares DIRP y sus interactores directos (véase el material suplementario, capítulo 5.3.7) muestra que la mayoría de estas interacciones físicas compartidas están citadas en la literatura o anotadas en bases de datos funcionalmente revisadas, aunque bastantes de estas interacciones permanecen aún sin publicar a la espera de un estudio funcional (véase la tabla S5 en el material suplementario, capítulo 5.3.7). El conjunto de interacciones compartidas publicadas constituye una validación positiva que apoya la hipótesis de una intercomunicación en las señales en red mediadas por los parálogos Ras pertenecientes al grupo DIRP.

Material suplementario

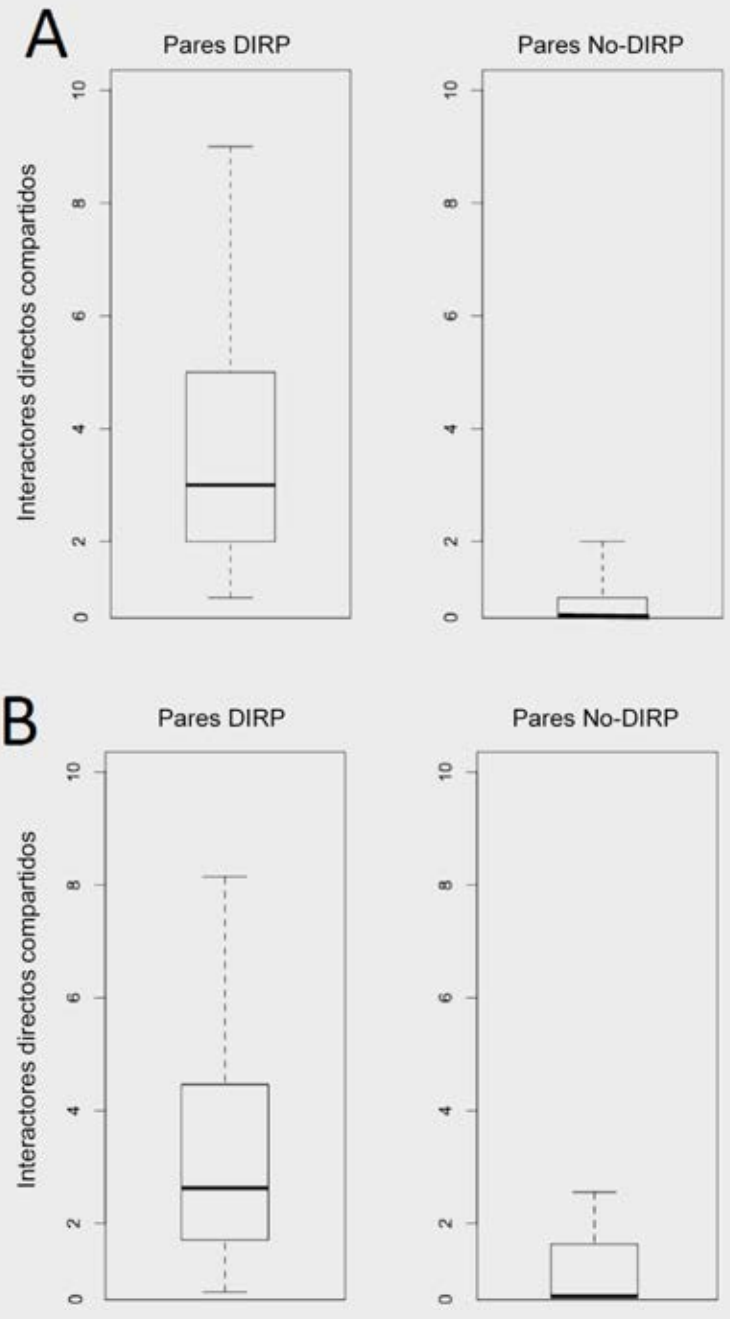


Figura 5.11: **Distribución del número de proteínas interactoras directas compartidas entre los parálogos Ras en los conjuntos de datos DIRP y No-DIRP.** Basado en CT (panel A) y DK (panel B).

5.3.3. Buscando posiciones conservadas en pares RAS divergentes pero interactuantes (DIRPs)

Con la finalidad de encontrar las posiciones específicamente conservadas dentro del conjunto de pares DIRP, todas las secuencias de las proteínas RAS fueron alineadas en un alineamiento múltiple de secuencias general (MSA). Posteriormente, para cada posición de aminoácidos, normalizamos su valor de conservación en el grupo positivo (DIRP) y el negativo (modelo aleatorio) comparándolos con la conservación, de las mismas posiciones, en el MSA completo (conjunto de datos de referencia) -véase la figura 5.8 y Material y métodos-. Esta normalización permitió identificar posiciones significativamente y específicamente conservadas en el conjunto DIRP comparadas con ambos conjuntos de datos (el aleatorio y el MSA de referencia completo). Haciendo uso de esta aproximación se seleccionaron un total de 22 posiciones (p-valor $\leq$ 0,01; puntos de corte superior e inferior en la figura 5.12) específicas del conjunto de datos DIRP. 21 de esas posiciones mostraron una mayor conservación en el grupo DIRP, mientras que únicamente 1 de las 22 posiciones manifestó una mayor variabilidad (menor conservación) dentro del conjunto de datos DIRP (posición R139 haciendo uso de HRas como secuencia de referencia en el alineamiento, véase la tabla S1 en el material suplementario -capítulo 5.3.7- y en el enlace: http://goo.gl/wcdiKV). La ausencia de pares de proteínas Ras similares en secuencia pero separadas en el interactoma (Área III en la figura 5.10) contrasta con la abundancia de pares Ras altamente divergentes y cercanos en la red (Área II en la figura 5.10). Esto sugiere que una proteína necesita acumular muchas mutaciones puntuales neutrales y adaptativas para poder tener nuevos compañeros de interacción, mientras que puede mantener su contexto de interacción a través de un número menor de posiciones conservadas clave.

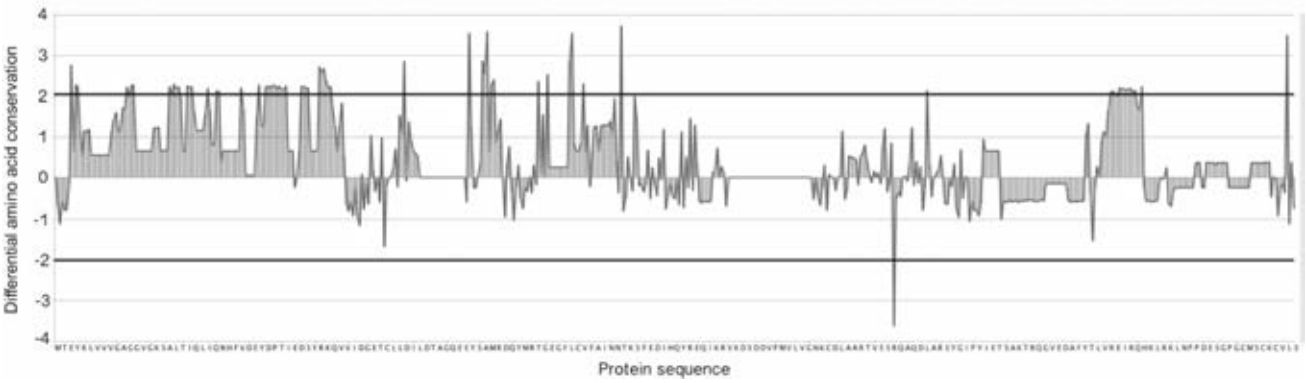


Figura 5.12: **Variación en la conservación de aminoácidos.** Diferencia de conservación en cada posición entre las secuencias del conjunto DIRP y el alineamiento general -MSA- (eje y). Los residuos utilizados como referencia de la secuencia peptídica corresponden a los aminoácidos de HRas en *Homo sapiens* (eje x). Un valor positivo en la variación de la conservación indica una posición con una mayor conservación en el conjunto de datos de los pares DIRP que en el set de referencia (MSA), y un valor negativo indica una posición con mayor variabilidad en el conjunto DIRP que en el alineamiento de referencia (MSA). Las líneas horizontales oscuras se corresponden con los puntos de corte asociados a p-valores $\leq 0{,}01$ calculados a partir del modelo aleatorio (control negativo).

5.3.4. Relación entre las posiciones conservadas en los DIRP y las regiones de unión de las proteínas Ras

Con el fin de investigar la relación entre las posiciones específicas de los DIRP y los sitios de unión de las proteínas Ras con otras proteínas, se recopilaron 28 complejos humanos de RAS de *Protein Data Bank* (PDB) [49] y se agruparon en 6 grupos estructurales (véase Material y métodos, capítulo 5.2.9). Tras esto se definieron las regiones de unión entre Ras y sus interactores en base al análisis de estos grupos estructurales (véase la tabla 5.3). De las 22 posiciones específicas identificadas en los DIRP en el paso anterior, 15 (68 %) están directamente implicadas en una o más regiones de unión y están localizadas en alguna de las regiones funcionales identificadas en las proteínas Ras (tabla 5.4, tabla 5.5 y figura 5.13). Otras 4 están rodeadas por 2 posiciones interactuantes consecutivas en la secuencia de aminoácidos. Considerando que estos últimos casos muy probablemente están también involucrados en las interacciones proteína-proteína de Ras, se puede

concluir que un 86 % de las posiciones específicas de los DIRP participa en las interacciones de Ras con otras proteínas (véase la tabla 5.4). Las restantes 3 no fueron relacionadas con ningún sitio de interacción conocido en el análisis. Estos resultados indican que las posiciones específicas de los DIRP son importantes para establecer interacciones entre Ras y su entorno en la red PPI y por lo tanto su conservación puede ser un factor importante a la hora de mantener a estos parálogos, distantes filogenéticamente, cercanos en el interactoma.

Functional Group	Complexes	Description	Positions	Num	Ratio
RasGef	1LFD	Interaction of Ras with RalGDS	**G12**, **Y32**, **D33**, **P34**, **I36**, E37, D38, S39, Y40, Q61, E62, E63, **Y64**, S65, **A66**, M67	7/16	43.7%
	1NVU 1NVX 1NVW 1NVV	Feedback activation by Ras. GTP of the Ras-specific nucleotide exchange factor SOS	S17, **T20**, I21, **Q22**, I24, N26, H27, D30, E31, **Y32**, **D33**, **P34**, **I36**, E37, D38, Y40, K42, Q43, V44, **D54**, I55, D57, **A59**, **G60**, Q61, E63, **Y64,** S65, **A66**, M67, D69, Q70, **Y71**, R73, R102, R149	12/36	33.3%
	1XD2	Autoinhibition in the Ras activator Son of sevenless: ternary Ras:SOS:Ras*GDP complex	**Q22**, I24, N26, H27, **D33, P34**, **I36**, E37, D38, K42, Q43, V44, L56, E63, **Y64**, **A66**, M67, Q70, **Y71**, R149	7/20	35.0%
	1BKD	The structural basis of the activation of Ras by Sos: H-Ras with SOS-1	S17, I21, **Y32**, **P34**, Y40, **D54**, I55, D57, **A59, G60**, Q61, E63, **Y64**, S65, **A66**, M67, D69, Q70, **Y71**, R73, R102	8/21	38.1%
RapGef	3CF6	Epac2 in complex with a cyclic AMP analogue and RAP1B	S17, **T20**, I21, H27, **Y32**, **P34**, E37, Y40, **D54**, I55, L56, D57, **A59**, **G60**, Q61, **Y64**, **A66**, M67, D69, Q70, **Y71**, Q99	9/22	40.9%
RasGap	1WQ1	The Ras-RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants	A11, **G12**, G13, I21, **Y32**, **D33, P34**, **I36**, E37, D38, S39, Y40, **G60**, Q61, E62, E63, **Y64**, K88, D92	7/19	36.8%
Antobodies (Cancer supressors)	2UZI	Tumour prevention by a single antibody domain targeting the interaction of signal transduction proteins with RAS	I21, V29, **D33, P34**, **I36**, E37, D38, Y40, Q61, **Y64**	4/10	40.0%
	2VH5	HRAS(G12V) - ANTI-RAS FV (DISULFIDE FREE MUTANT) COMPLEX	I21, V29, **D33, P34**, **I36**, E37, D38, Y40, D57, Q 61, **Y64**	4/11	36.4%
	3DDC	Ras effector interaction between tumour suppressor NORE1A and Ras switch II	I24, Q25, **I36**, D38, Y40, **Y64**, M67	2/7	28.6%
Ras Binding Domain & PI3K	1HE8	Ras binding to its effector phosphoinositide 3-kinase gamma	I21, **D33**, **I36**, E37, D38, S39, Y40	2/7	28.6%
	3KUC 1GUA	Complex of Rap1A(E30D/K31E) GDP with RafRBD(A85K/N71R) Ras/Rap effector specificity determined by charge reversal	I21, **D33**, **I36**, E37, D38, S39, Y40, R41	2/8	25.0%
	3KUD	What makes Ras an efficient molecular switch: Ras-GDP interactions with mutants of Raf	I21, E37, D38, S39, Y40, R41	0/6	0.0%
	1K8R	Ras-Byr2RBD complex: structural basis for Ras effector recognition	**I36**, E37, D38, S39, Y40, R41, **D54**	2/7	28.6%

Functional Group	Complexes	Description	Positions	Num	Ratio
	1C1Y	c-Raf1 in complex with Rap1A and a GTP analogue	I21, **I36**, E37, D38, S39, Y40, R41	1/7	14.3%
Other cases	1ZC3 1ZC4	Ral-binding domain of Exo84 in complex with the active RalA	D47, G48, E49, T50, C51, L52, M67, G75, F78, V81, **F82**	1/11	9.1%
	2A9K 2A78	C3bot-NAD-RalA complex: Ral-A and Mono-ADP-ribosyltransferase C3 C3bot-RalA complex	**T20**, I21, **Q22**, L23, D69, **Y71**, M72, G75, L79, A83, **V103**, S106, D107, P110	4/14	28.6%
	2C5L	PLC epsilon Ras association domain with HRas	I24, Q25, **I36**, D38, S39, Y40, D47, S127, Q131, A134, Y141, I142, E143, D154, R161, R164, Q165	1/17	5.9%
	4DXA	Rap1 in complex with KRIT1	Q25, H27, **I36**, E37, D38, S39, Y40, Q43, M67	1/9	11.1%
	3T5G	Rheb in complex with PDE6D	T2, D57, G178, P179, G180	0/5	0.0%
	2BOV	recognition of an ADP-ribosylating Clostridium botulinum C3 exoenzyme by RalA GTPase	**I139**, P140, E143	1/3	33.3%
	1UAD	Interaction between RalA and Sec5, a subunit of the sec6/8 complex	**I36**, G48, E49, T50, C51	1/5	20.0%

Tabla 5.3: **Mapeo de las posiciones específicas en los DIRP en los sitios de unión de los complejos de Ras en *Homo sapiens*.** De izquierda a derecha: Etiquetas de los grupos funcionales para los conjuntos de complejos; códigos de identificación en PDB de los complejos tridimensionales de Ras agrupados con r.m.s.d. $< 1{,}0$; descripción de los complejos 3D de Ras; posiciones involucradas en los sitios de unión de los complejos Ras (aquellas que coinciden con las posiciones específicas en los DIRP se muestran en negrita); número (Num.) y porcentaje (Ratio) de las posiciones específicas en los DIRP mapeadas en relación al total de posiciones existentes en los sitios de unión. La numeración de las posiciones está referenciada a la secuencia de la proteína HRas en humanos.

Position	Number of matches	Complexes
I36	14	1NVU, 1XD2, 1LFD, 1WQ1, 2UZI, 2VHS, 3DDC, 1H8E, 3KUC, 1K8R, 1C1Y, 2C5L, 4DXA, 1UAD
Y64	9	1NVU, 1XD2, 1BKD, 1LFD, 3CF6, 1WQ1, ZUZI, 2VHS, 3DDC
D33	8	1LFD, 1NVU, 2UZI, 2VH5, 1HE8, 1XD2, 1WQ1, 3KUC
P34	8	1LFD, 1NVU, 2UZI, 2VH5, 1XD2, 1WQ1, 1BKD, 3CF6
Y32	5	1NVU, 1BKD, 1LFD, 3CF6, 1WQ1
A66	5	1NVU, 1XD2, 1BKD, 1LFD, 3CF6
Y71	5	1NVU, 1XD2, 1BKD, 3CF6, 1ZC3
D54	4	1NVU, 1BKD, 3CF6, 1K8R
G60	4	1NVU, 1BKD, 3CF6, 1WQ1
T20	3	1NVU, 3CF6, 2A9K
A59	3	1NVU, 1BKD, 3CF6
Q22	3	1NVU, 1XD2, 2A9K
G12	2	1LFD, 1WQ1
V103	1	2A9K
I139	1	2BOV
T35	0	Between interacting positions 34 & 36 in several complexes
R68	0	Between interacting positions 67 & 69 in several complexes
T58	0	Between interacting positions 57 & 59 in some complexes
F28	0	Between interacting positions 27 & 29 in some complexes
G77	0	
E153	0	
C186	0	

Tabla 5.4: **Listado ordenado de las posiciones específicas en los DIRP basado en su nivel de implicación en los sitios de unión de Ras.** La primera columna muestra la posición y el aminoácido de acuerdo a la secuencia de HRas en *Homo sapiens*. La segunda columna indica el número de sitios de unión en complejos 3D en los cuales la posición se encuentra directamente implicada. La tercera columna contiene los códigos de identificación PDB para los complejos que están relacionados con cada posición -o anotaciones en caso de relaciones indirectas a sitios de unión-.

Functional Regions	Positions	Ratio
Switch I (Effectors binding site)	Y32, D33, P34, T35, I36	23%
Switch II	G60, Y64, A66, R68, Y71	23%
C-terminal hyper variable region	C186	5%
Nucleotide (GDP/GTP) binding site	G12, F28, T35, T58, A59, G60	27%
Innert regions	T20, Q22, D54, G77, V103, I139, E153	32%

Tabla 5.5: **Posiciones específicas en los DIRP agrupadas en regiones funcionales de Ras.** La primera columna muestra un listado de las diferentes regiones funcionales en las proteínas Ras. La segunda columna indica las posiciones y aminoácidos de los DIRP de acuerdo a la secuencia de HRas en *Homo sapiens*. La tercera columna muestra el porcentaje (Ratio) de posiciones específicas de los DIRP en cada región funcional.

Las posiciones específicas de los DIRP constituyen un gran porcentaje ($\approx$ 38 %) de la región de unión de Ras con los efectores *Guanine Exchange Factor (GEF)* (véase la tabla 5.3), tales como SOS (Ras GEF), Epac2 (Rap GEF), RalGDS (Ras GEF) y *GTPase Activating Protein* (GAP). Las posiciones aminoacídicas de los DIRP seleccionadas son también importantes para las regiones de interacción de Ras con anticuerpos supresores de tumores ($\approx$ 35 %) y, en menor medida, para el Dominio de Unión a Ras (*Ras Binding Domain* -RBD-) de diferentes efectores activados por la señal de Ras ($\approx$ 19 %), tales como *phosphoinositide 3-kinase*, *Raf*, *Byr2* y *c-Raf1*. Adicionalmente, varias posiciones específicas de los DIRP mapean con residuos frecuentemente mutados en cáncer (véase la figura 5.14 y la tabla S6 en el material suplementario -capítulo 5.3.7- y en el enlace: goo.gl/eXPW9i). Esto es particularmente evidente para residuos tales como G12, que conjuntamente con G13 y Q16 concentran el 97 % de las mutaciones oncogénicas de RAS [161].

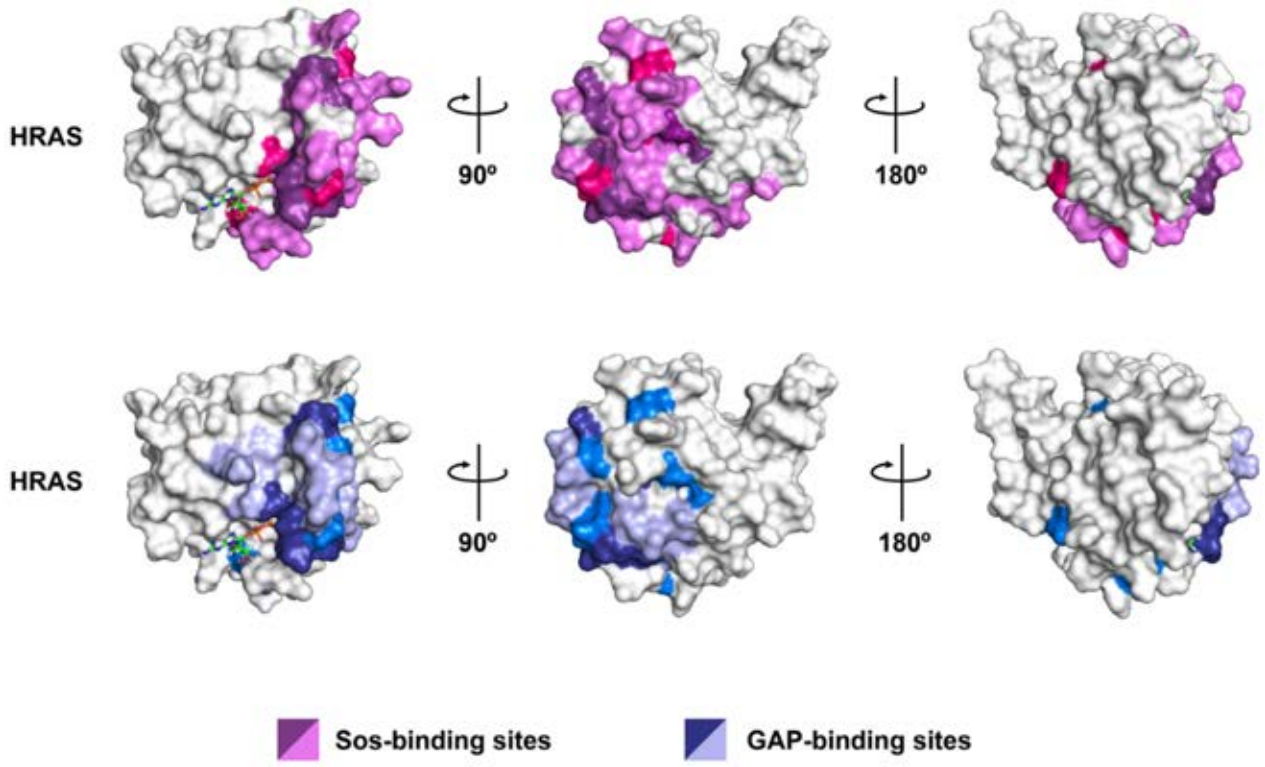


Figura 5.13: **Distribución espacial de todas las posiciones específicas de los DIRP en la proteína HRas.** Los residuos relevantes son mapeados en un modelo de superficie de la estructura 3D del parálogo humano de HRas (pdb# 1aa9). Fila superior: los residuos implicados en la interacción de HRas con *GEF Sos* se muestran en morado claro mientras que aquellas posiciones específicas de los DIRP involucradas en la interacción se destacan en morado oscuro. Las posiciones específicas de los DIRP no implicadas en la interacción HRas-Sos están coloreadas en rosa. Fila inferior: los residuos involucrados en la interacción de HRas con GAP se muestran en azul claro mientras que aquellas posiciones específicas de los DIRP involucradas en la interacción se destacan en azul oscuro. Las posiciones específicas de los DIRP no implicadas en la interacción HRas-GAP están coloreadas en azul marino. Nótese que con esta aproximación se pueden identificar las posiciones específicas de los DIRP que coinciden con las zonas de interacción y las que no. Para mayor claridad se presentan, en cada caso, 3 rotaciones tridimensionales diferentes de la proteína HRas.

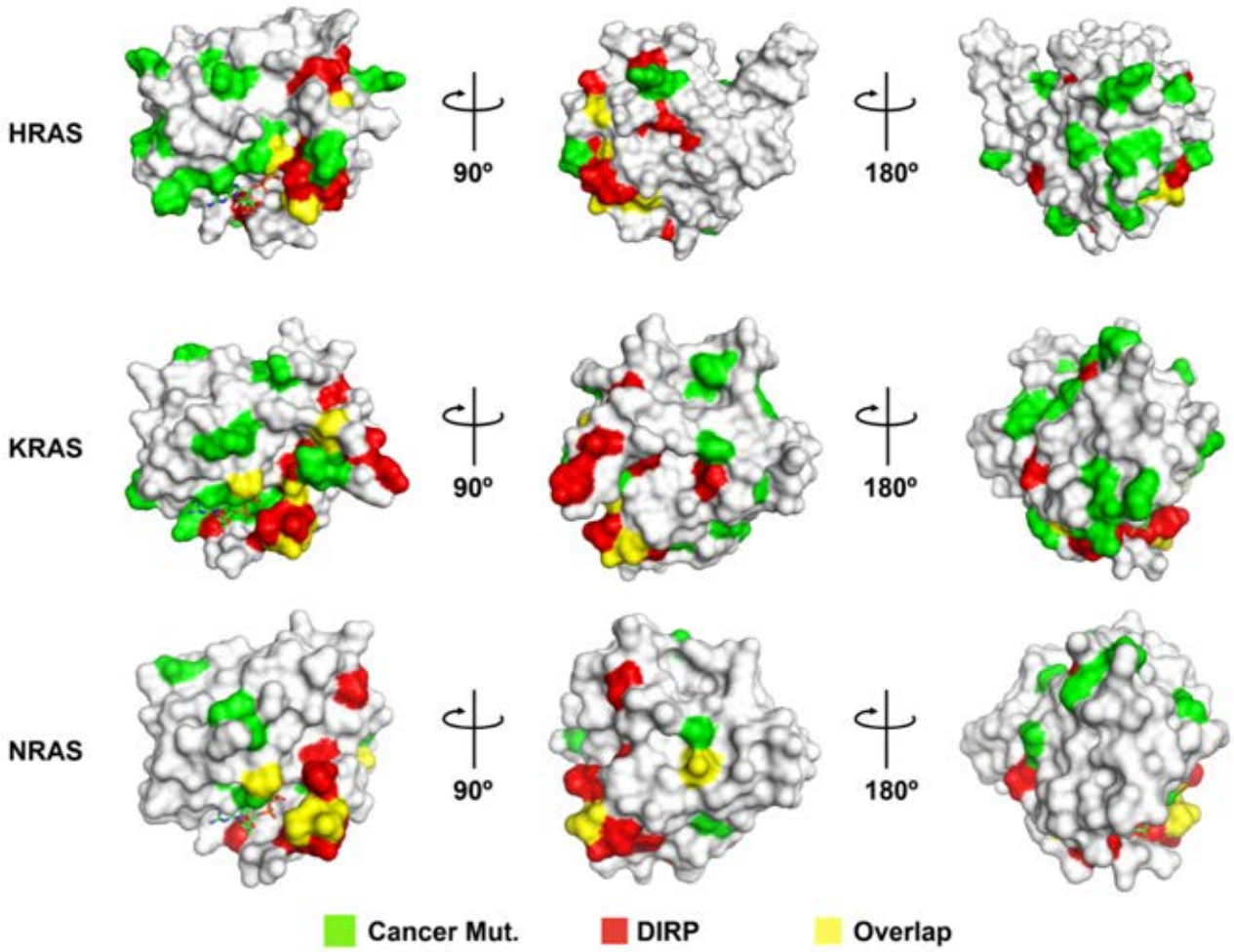


Figura 5.14: **Solapamiento de las posiciones específicas de los DIRP en zonas frecuentemente mutadas en cáncer o sus cercanías.** Se muestra la superficie de los modelos estructurales en 3D de HRas (pdb# 1aa9), KRas (pdb# 4epv) y NRas (pdb# 3con; la única estructura tridimensional de NRas disponible en PDB, la cual carece de los residuos 61-71), los 3 parálogos humanos RAS más frecuentemente mutados en cáncer. Las mutaciones que actualmente se contemplan en el catálogo TCGA (Cancer Mut.) han sido coloreadas en verde y las posiciones específicas de los DIRP en rojo. Las posiciones solapantes (*i.e.* posiciones específicas de los DIRP que se corresponden con residuos frecuentemente mutados en cáncer) se muestran en color amarillo. Para mayor claridad se presentan, en cada caso, 3 rotaciones tridimensionales diferentes para cada proteína.

Otros complejos 3D de Ras muestran una muy baja implicación de estas posiciones específicas de los DIRP en sus regiones de unión en Ras. Por ejemplo, el dominio de asociación a Ras de la proteína *PLC epsilon* únicamente solapa con 1 posición de un total de 17 (véase la tabla 5.3). Los resultados en este caso sugieren una baja influencia de las posiciones específicas de los DIRP en la señal mediada por este domino. Únicamente 2 complejos no tienen coincidencia alguna con las posiciones de los DIRP: el complejo Ras con una proteína Raf mutada y la interacción de Rheb (proteína homóloga a Ras) con la proteína PDEδ, un supuesto factor de solubilización para varias proteínas preniladas de la subfamilia Ras [162].

Tal como se ha mencionado, 3 posiciones de los DIRP (G77, E153 y C186) no coinciden con ninguna región de unión, algo que puede deberse a la ausencia de algunos complejos Ras no registrados todavía en PDB. Específicamente, es sabido que la posición 186 es un residuo que se corresponde con un aminoácido Cys conservado, localizado en una región C-terminal desestructurada y altamente variable de la proteína Ras. Funcionalmente, este residuo está involucrado en el reconocimiento específico de la molécula y su transporte a la superficie interna de la membrana plasmática, sin el cual la proteína está inactiva [163–165]. El alto grado de conservación en esta posición podría sugerir una colocalización de los pares de proteínas DIRP en la membrana interna.

5.3.5. Discusión

En este trabajo se llevó a cabo un análisis exhaustivo de la relación entre la filogenia de las proteínas RAS y su localización en la red de interacciones. A esto le siguieron análisis de secuencia y estructurales de las posiciones conservadas en los pares DIRP en los sitios de unión de RAS con sus efectores. Los análisis de secuencia de estas proteínas divergentes pero interactuantes identificaron esas posiciones clave, las cuales mapeaban en las regiones de unión 3D que en Ras mediaban las interacciones con muchos de sus efectores. Estos resultados sustentan la idea de que estas posiciones conservadas determinan qué pares DIRP están próximos en el interactoma, *i.e.* compartiendo contextos de interacción similares.

La destacada relación de las posiciones específicas de los DIRP con los sitios de unión en Ras sugiere que mutaciones puntuales de estas posiciones en células somáticas podrían resultar en un recableado de la red Ras, llevando a estados patológicos [166], particularmente en aquellas mutaciones que afectan al interruptor de regulación *on/off*. Las mutaciones en las proteínas Ras pueden llevar a un estado permanentemente activado de proliferación celular o a una alteración de la red de interacciones de Ras que derive en desarrollo tumoral [167]. Adicionalmente, se conoce que el cambio de únicamente un par de residuos clave entre las proteínas parálogas Ras y Ral produce el intercambio de especificidad entre sus efectores naturales [168, 169]. Uno de estos residuos intercambiados entre Ras y Ral es el I36 (según coordenadas de la secuencia HRas), el cual se corresponde con la posición específica de los DIRP involucrada en el mayor número de sitios de unión de complejos Ras (véase la tabla 5.4). Otras posiciones específicas de los DIRP coinciden con conocidas regiones de unión supresoras de tumores en Ras, sugiriendo que una investigación más a fondo de las posiciones conservadas en los DIRP podría inspirar nuevas aproximaciones anti-tumorales. La metodología que se describe en este trabajo podría ser extendida al estudio de otras familias de proteínas, haciendo uso del mismo procedimiento.

El hecho de que un alto número de parálogos Ras distantes compartan su contexto de compañeros de interacción en la red mediante la conservación de unas pocas posiciones clave, hace plausible la hipótesis de una evolución convergente en la red de interacciones de Ras. No obstante, el modelo filogenético observado en este trabajo muestra que las proteínas Ras se alejan en el interactoma humano mediante la acumulación de muchas mutaciones puntuales neutrales y adaptativas en un gran proceso de divergencia de secuencias, ya que no se observan parálogos Ras cercanos en el árbol filogenético y distantes en el interactoma.

De todos modos, también es posible que la convergencia en la red de interacciones se produzca cuando las secuencias Ras divergen. Ciertamente, el estudio del papel potencial de la evolución convergente en la configuración de la red de señalización de Ras es un tema clave que merece un análisis filogenético más profundo.

A pesar de los intensos esfuerzos en el campo, tanto en la investigación básica como en la aplicada, a lo largo de los últimos 30 años, todos los intentos de desarrollar un inhibidor de RAS efectivo han fracasado y en consecuencia las proteínas RAS han sido históricamente consideradas como farmacológicamente intratables [151, 161, 170]. La mayoría de los estudios han tratado o bien de bloquear la farnesilación de RAS para impedir su translocación a la membrana plasmática o bien de interferir en la unión del nucleótido GTP, impidiendo de estas maneras la función de RAS. Los inhibidores de la farnesiltransferasa de RAS han fallado básicamente porque las células pueden hacer uso de rutas alternativas para añadir modificaciones postraduccionales a las proteínas RAS. Por otro lado, las GTPasas RAS se unen a nucleótidos con afinidades muy altas (del orden picomolar), lo que hace muy difícil a un inhibidor competir en afinidad con los nucleótidos intracelulares, los cuales se encuentran en el rango milimolar [171]. Más recientemente, no obstante, multitud de grupos de investigación han contribuido con nuevas estructuras 3D que muestran a las GTPasas RAS en conformaciones previamente desconocidas [172]. Este conjunto de datos, junto con nuevos modelos computacionales dinámicos de la activación de RAS y una nueva metodología basada en una combinación de ingeniería de proteínas y síntesis orgánica, *i.e.* genética química [173, 174], han revelado sitios de unión transitorios en las estructuras 3D de las proteínas RAS que pueden ser usados como diana por pequeñas moléculas inhibidoras, lo que conduce a un renovado interés en las proteínas RAS como dianas de nuevos fármacos [170]. Siguiendo los enfoques de modelado computacional, se han diseñado nuevas moléculas para inhibir la función de RAS y RAL. No se han descrito inhibidores de RAP hasta la fecha. 2 péptidos ortostéricos, HBS3 [175] y SAH-SOS1 [176], afectan de manera eficaz a las interacciones Ras-GEF mimetizando la hélice αH de SOS1 posicionada entre las regiones *switch I* y *switch II* en Ras, que implica a los residuos L6, G15, L56, D57, E63, Y64, R73, T74 y Q99 en KRAS. La mayoría de estos residuos están próximos (o son idénticos, *e.g.* Y64) a algunos de los residuos específicamente conservados en los pares DIRP identificados en este trabajo. Además, varios grupos han tenido éxito recientemente en el uso de pequeñas moléculas inhibidoras de las interacciones específicas del complejo Ras-GEF: mediante el análisis de diferentes conformaciones de RAS, se encontraron nuevos sitios de unión potencialmente tratables con este tipo de inhibidores, implicando a los residuos: K5, L6, V7, D54

(conservado en los DIRP), I55, L56, Y71 (conservado en los DIRP) y T74 [177, 178]. Adicionalmente, haciendo uso de una criba basada en RMN, Sun *et al.* [179] identificaron un sitio de unión hidrofóbico localizado entre la hélice α2 del *switch II* (residuos 60-70, región en la que encontramos los residuos conservados en los DIRP: 60, 64, 66 y 68) y la lámina β central de KRas-G12D, regiones donde es posible acoplar un conjunto de inhibidores de moléculas pequeñas, bloqueando la interacción con su efector SOS-GEF.

También se han identificado inhibidores que dificultan la unión de RAL a sus efectores GEF aguas arriba en la ruta de señalización, mediante la aplicación de cribado virtual sobre la estructura de esta proteína. Siguiendo esta metodología se identificaron 3 nuevos compuestos (RBC6, RBC8 y RCB10) capaces de interactuar con un sitio de unión al efector GEF, adyacentes al *switch II* (residuos 70-77) y a la hélice α2 (residuos 78-85) de RALA [180]. Mediante el uso de acoplamiento molecular, se predijo que los residuos implicados en la interacción en HRAS eran los correspondientes a las posiciones T58, G60, R68, Y71 y M72, todos los cuales (excepto M72) se identificaron como residuos conservados en los DIRP en el análisis llevado a cabo en este trabajo. Curiosamente, posiciones equivalentes a G10, A11 y Q95 en HRAS fueron predichas como mediadoras en la unión de inhibidores RBC a RALA interfiriendo en la interacción de RALA con GEF, estando estos 3 residuos cerca de otras posiciones conservadas en los DIRP, *i.e.* R103 y el G12 catalítico. Por lo tanto, independientemente de su naturaleza química (péptidos o moléculas pequeñas) el nuevo conjunto de compuestos inhibidores diseñados para bloquear las interacciones proteína-proteína en la red de la familia Ras comparte varios residuos diana críticos que son idénticos a algunas posiciones conservadas en el grupo DIRP identificadas en este estudio.

En contraste con las proteínas prototípicas RAS, las mutaciones en las proteínas RAL o RAP son poco frecuentes e irrelevantes en cáncer (véanse las tablas S6 y S7 en el material suplementario -capítulo 5.3.7- y en los enlaces: http://goo.gl/eXPW9i y goo.gl/4tD9i3) [181]. No obstante RALA y RALB se encuentran sobreexpresadas en un conjunto de tumores, siendo los más destacables el NSCLC (*Non-small-cell lung carcinoma*, o carcinoma pulmonar no microcítico) y el melanoma [154, 155, 182]. En base a lo expuesto, parece poco probable que se produzca una reconfiguración

de la red de señalización Ras por cambio de pares interaccionantes como consecuencia de mutaciones puntuales en residuos conservados en los DIRP. La razón es que las mutaciones oncogénicas puntuales solo se han encontrado en las proteínas HRAS, KRAS y NRAS, y con las posiciones G12 y Q61 representando la gran mayoría de los casos (97 % en HRAS, 99 % en KRAS) (véanse las tablas S6 y S7 en el material suplementario -capítulo 5.3.7- y en los enlaces: http://goo.gl/eXPW9i y goo.gl/4tD9i3) [161]. Por otro lado, puede producirse una reconfiguración de las interacciones en la red debido a cambios en la expresión de proteínas en el contexto de las proteínas RAL, pues aunque la expresión alterada de isoformas RAS no es una característica común en el cáncer (véase la figura S1 del material sumplementario, capítulo 5.3.7) estos cambios sí se relacionan con algunas RASopatías [161, 183]. Sin embargo, los resultados presentados aquí, *i.e.* la identificación de residuos conservados en los DIRP coincidentes con posiciones ocupadas por nuevos inhibidores de interacciones proteína-proteína de las GTPasas RAS, sugieren que este nuevo grupo de inhibidores podría no ser tan específico como se esperaba inicialmente. Esto es particularmente importante ya que aún falta información sobre su eficacia *in vivo*. Se ha demostrado que los ortopéptidos HBS3, SAH-SOS1 y el compuesto DCAI reducen los niveles de Ras-GTP y, en algunos casos, inhiben la activación de ERK en células cultivadas [175–177], pero no se ha informado todavía sobre experimentos *in vivo*. Por el contrario, el RBC8 y algunos inhibidores relacionados con la interacción RAL-GEF se han probado en ratones xenoinjertados con tumores H2122 (pulmón), donde pudieron reducir el crecimiento tumoral de una manera dependiente de la dosis [180]. Los inhibidores de Ras podrían usarse en combinación con otros inhibidores de la ruta ERK ya que, por ejemplo, el bloqueo de la actividad MEK sola no es efectiva en la inhibición de tumores dependientes de Ras. Además, la inhibición oncogénica de BRAF (V600E) puede provocar paradójicamente la activación de la vía [170, 184, 185]. En cualquier caso, a la luz del renovado interés en las GTPasas RAS como dianas farmacológicas en cáncer [31], parece que la identificación de residuos conservados en los DIRP debería ser una herramienta valiosa para ayudar en la evaluación de las posibles inespecificidades de los nuevos inhibidores de Ras.

Las redes PPI utilizadas en este estudio se basan en interacciones físicas entre proteínas recolectadas de diferentes fuentes de información, incluyendo experimentos *in vitro*. Estos últimos no consideran todas las regulaciones temporales o espaciales de la expresión génica (*e.g.* las barreras de los distintos compartimentos celulares), las cuales podrían evitar que algunas de las interacciones ocurran *in vivo*.

Los resultados de este estudio añaden una perspectiva novedosa al modelo generalmente aceptado según el cual los genes parálogos filogenéticamente cercanos tienen interacciones similares que divergen a lo largo del tiempo junto con la divergencia de sus secuencias [32–34]. Aunque la especificidad de las interacciones proteína-proteína es el resultado de una compleja combinación de factores, este trabajo sugiere que existe un conjunto de posiciones aminoacídicas clave que son altamente relevantes para la especificidad de las interacciones. Estas posiciones podrían explicar el motivo por el cual proteínas Ras divergentes comparten contextos de interacción similares, incrementando la probabilidad de que exista intercomunicación en las señales mediadas por estas proteínas Ras. Encontrar compuestos que tengan como objetivo estas regiones funcionalmente solapantes de los pares DIRP podría ayudar en el diseño de nuevas estrategias terapéuticas.

5.3.6. Publicación: *Exploring the interactions of the RAS family in the human protein network and their potential implications in RAS-directed therapies. Oncotarget 2016.*

El trabajo de investigación descrito en este capítulo dio como fruto la publicación del artículo que aquí se adjunta y que sirve como aval de esta Tesis Doctoral.

5.3.7. Material Suplementario

A continuación se presentan los resultados adicionales del estudio, publicados como material suplementario del artículo *Exploring the interactions of the RAS family in the human protein network and their potential implications in RAS-directed therapies. Oncotarget 2016*.

Capítulo 6

Asociaciones fenotipo-*loci* en redes de pacientes con trastornos genéticos raros: aplicación en la asistencia al diagnóstico de nuevos casos clínicos

No es sólo que el universo sea más extraño de lo que pensamos; es que es más extraño de lo que podemos pensar.

Werner K. Heisenberg

6.1. Introducción

Décadas de avances en tecnologías relacionadas con la genómica están incrementando significativamente la precisión en el campo del diagnóstico genético. Actualmente está ampliamente aceptado el hecho de que una detallada descripción fenotípica de los pacientes [51], junto con su caracterización genotípica, tiene un enorme potencial a la hora de agilizar la identificación de nuevos síndromes, con sus ventajas asociadas en la determinación del pronóstico o tratamiento, así como en la comprensión de las bases genéticas de las propias enfermedades humanas [10, 52, 53]. De hecho, el diagnóstico de los trastornos genéticos en humanos por parte de un genetista clínico es mucho más sencillo cuando los fenotipos relacionados con un síndrome han sido reconocidos previamente o cuando existe la suficiente cantidad de datos clínicos extraídos de un elevado número de pacientes que comparten el mismo (o al menos similar) reordenamiento genómico y conjunto de fenotipos en enfermedades genéticas aún no reconocidas [62]. Sin embargo, realizar un diagnóstico con precisión en muchos casos de trastornos genéticos se vuelve complicado cuando los pacientes poseen perfiles fenotípicos complejos [54], cuando varios síndromes cromosómicos comparten características clínicas entre ellos, o cuando aberraciones genéticas raras afectan a un número de pacientes extremadamente bajo, tal como ocurre en el caso de las enfermedades raras. Por lo tanto, entre los retos más importantes a los que actualmente se enfrentan los profesionales clínicos se encuentran la interpretación y clasificación de nuevas y/o extremadamente raras variantes genéticas y la comprensión de sus consecuencias fenotípicas en patologías complejas. La aproximación '*partiendo del genotipo*', en la cual los pacientes son clasificados inicialmente por presentar reordenamientos genómicos similares y posteriormente por compartir fenotipos, ha demostrado ser de gran utilidad en la caracterización de la creciente lista de síndromes asociados a microdeleciones y microduplicaciones. De hecho, haciendo uso de esta aproximación, nuevos síndromes asociados a microdeleciones o microduplicaciones han sido descritos recientemente [55, 56].

Array-Comparative Genomic Hybridization (aCGH) y *Single Nucleotide Polymorphisms arrays (SNParrays)*, junto con *Next Generation Sequencing* (NGS), son las principales tecnologías utili-

zadas para detectar *copy number variations* (CNVs) [101]. Las CNVs son variaciones estructurales del genoma que van desde pequeñas mutaciones (1kb) a grandes cambios estructurales (millones de nucleótidos). Estas variaciones pueden corresponder a deleciones, duplicaciones o translocaciones encontradas en distintas regiones genéticas, bien heredadas o bien producidas espontáneamente en el propio individuo (*de novo*), las cuales en algunos casos pueden producir enfermedad [7]. Aunque también existen CNVs en individuos sanos -estas representan alrededor del 4,8-9,5 % de la variación del genoma humano [101]- como variación natural entre genomas en la población. De todos modos, las CNVs heredadas o *de novo* pueden ser la causa de muchos trastornos (tales como esquizofrenia, enfermedad de Crohn o autismo) y su identificación y análisis son utilizados para el diagnóstico y caracterización de diversos síndromes cromosómicos [57–59]. En algunos laboratorios, los *microarrays* son una tecnología heredada que será reemplazada por las conocidas como NGS (*Next Generation Sequencing*). Sin embargo, el todavía creciente número de pacientes genotipados mediante plataformas aCGH y SNP *array* sugiere un uso generalizado de esta tecnología para el descubrimiento de nuevas entidades en genética clínica. De hecho, bases de datos públicas tales como DECIPHER [7, 62, 63] muestran una significativa cantidad de datos procedentes de tecnologías aCGH y SNP *array* en los últimos años.

Hoy en día, la identificación completa de las consecuencias fenotípicas de una CNV dada sigue siendo un reto. Es necesario tener en cuenta un amplio número de potenciales mecanismos moleculares y genéticos para determinar su relación con los fenotipos del paciente: *imprinting*, desenmascaramiento de una mutación recesiva en el otro alelo, disfunción de elementos regulatorios o herencias complejas -interacciones entre más de una mutación- (véase el capítulo 1.4.4). Incluso una CNV clasificada como 'normal' puede tener efectos patogénicos en individuos con perfiles genéticos diferentes [62]. En este sentido, la integración y comparación a gran escala de fenotipos y genotipos de pacientes con enfermedades raras es esencial para poder alcanzar un diagnóstico, además de ser crucial para avanzar en la caracterización de las regiones genéticas y los mecanismos moleculares que controlan la expresión fenotípica en dichos pacientes.

Con la intención de ayudar en la caracterización de las relaciones moleculares entre diferentes fenotipos patológicos y microvariaciones, este trabajo se centró en la aplicación de principios de la medicina de redes (*Network Medicine*) [3–5, 60, 61]. Para ello, se implementó una aproximación computacional basada en redes *tripartitas* construidas a partir de las siguientes capas: i) variantes genéticas (CNVs), ii) pacientes y iii) fenotipos. Se hizo uso de la información disponible en la base de datos DECIPHER -un repositorio global de datos clínicos de pacientes- como recurso para un análisis sistemático que permitiera la identificación de las CNVs potencialmente patogénicas [62]. DECIPHER es una valiosa fuente de información que ofrece los registros fenotípicos y genotípicos de un considerable número de pacientes con trastornos genéticos de baja prevalencia, recopilados de más de doscientas instituciones clínicas a lo largo del mundo [7, 62, 63]. Se construyeron redes *tripartitas* con miles de asociaciones fenotipo-paciente-genotipo haciendo uso de los datos de 10.324 pacientes de la base de datos DECIPHER. La mayoría de los pacientes con CNVs *de novo* en la base de datos DECIPHER están relacionados con trastornos pediátricos asociados a retrasos en el desarrollo, retraso mental o anomalías estructurales congénitas [54, 64]. Las redes generadas incluyen 14.227 CNVs, detectadas mediante técnicas de *microarrays*, las cuales han demostrado ser eficientes a la hora de identificar nuevos síndromes cromosómicos en numerosos casos en los últimos años [55, 62, 64]. Los pacientes de DECIPHER son además anotados bajo un grupo controlado de términos (fenotipos patológicos) que provienen de la *Human Phenotype Ontology* (HPO) [10], incluyendo en DECIPHER un total de 2.583 términos diferentes. Esta anotación de fenotipos a través de una ontología formal y normalizada, como HPO, permite el análisis y la comparación sistemática de sintomatologías entre pacientes.

Finalmente, con el propósito de estudiar las relaciones fenotipo-genotipo en este conjunto de datos, se explotaron las asociaciones de la red *tripartita* construida *ad hoc*, analizando el subgrupo de pacientes con CNVs *de novo* en DECIPHER (incluyendo la información fenotípica) y se identificaron asociaciones estadísticamente significativas entre regiones mutadas y fenotipos patológicos. Estas asociaciones fenotipo-*locus* han sido tenidas en cuenta a la hora de evaluar el potencial de es-

ta aproximación basada en redes para la asistencia al diagnóstico de nuevos casos no caracterizados en la rutina clínica.

Esta aproximación muestra el potencial de integrar información genotípica y fenotípica de miles de pacientes para la identificación de nuevos patrones genotipo-fenotipo en el estudio de casos raros y aislados en los que escasea la información con la que compararlos.

6.2. Material y métodos

1.- Fuente de datos para la construcción de las redes

Se hizo uso de las CNVs *de novo* de los pacientes de DECIPHER con trastornos genéticos raros y anotados con términos HPO (versión 2014-05-08, mapeada al genoma de referencia hg19) a través de un acuerdo de acceso a la información con el consorcio propietario de la base de datos. Todos los datos sobre fenotipos y genotipos pertenecen a pacientes que han proporcionado su consentimiento informado a compartir estos registros de manera anónima. Esta información incluye el conjunto de términos HPO anotados para cada paciente y sus respectivas CNVs asociadas. La subred de deleciones incluye 2.436 CNVs *de novo* de 2.301 pacientes y 1.795 fenotipos HPO. La subred de duplicaciones está formada por 1.114 CNVs *de novo* de 1.013 pacientes, incluyendo 1.160 términos HPO. Para cada paciente, DECIPHER selecciona únicamente las CNVs potencialmente patológicas, eliminando aquellas observadas en las poblaciones control. La versión 2014 de DECIPHER se utilizó para construir las redes *tripartitas* y validar las hipótesis, debido a que esta versión no incluye los datos de los pacientes usados como *nuevos casos clínicos* en la fase de validación (véase la siguiente sección), cosa que sí ocurre en versiones más recientes, al haber sido incluidos éstos en la base de datos DECIPHER con posterioridad.

Se analizaron 2 tipos diferentes de relaciones: i) pacientes y genotipos (mediante las CNVs) y ii) pacientes y fenotipos (a través de los términos HPO). Se dividieron las CNVs en deleciones y duplicaciones, debido al hecho de que pueden tener efectos diferentes cuando afectan a la misma región. Por ejemplo, en los síndromes de microdeleción/microduplicación 19p13.3 y 19p13.13 [55, 66], las microdeleciones derivan en macrocefalia mientras que las microduplicaciones en microcefalia. Adicionalmente, el estudio se ha centrado únicamente en las mutaciones *de novo*, por ser aquellas que con mayor probabilidad están asociadas a fenotipos patológicos y corresponden con mayores reordenamientos genómicos [65].

2.- Casos clínicos usados para validar el método

Con la finalidad de probar la eficacia de este método, basado en redes, para la asistencia en el diagnóstico clínico de pacientes con ganancias o pérdidas en su genoma, se utilizaron 2 cohortes de pacientes proporcionados por el INGEMM (Instituto de Genética Médica y Molecular, Hospital La Paz, Madrid, España). Los datos de los pacientes fueron sometidos a un estricto control ético, consistente en la firma de formularios de consentimiento por parte de los propios pacientes o sus tutores legales. Los datos se obtuvieron en el mismo formato que los de DECIPHER: un conjunto anonimizado de CNVs y sus correspondientes términos HPO anotados por paciente. Las investigaciones clínicas fueron llevadas a cabo de acuerdo a las directrices de la Declaración de Helsinki [186]. Este grupo de pacientes fue utilizado en este trabajo como prueba de concepto con la cual validar la metodología aquí desarrollada y comprobar si las asociaciones que encontramos en las redes de DECIPHER podrían ayudar en la identificación de los fenotipos asociados a los casos de nuevas CNVs patológicas del grupo de pacientes del INGEMM. Para llevarlo a cabo, primero se aplicó la aproximación en red a la base de datos DECIPHER (versión 2014-05-08), versión que no incluye a los pacientes de INGEMM utilizados para la posterior validación del método. Tras esto y en base a los pesos de las asociaciones genotipo-fenotipo encontradas en la red de pacientes de DECIPHER, se realizó una prelación de las asociaciones putativas fenotipo-CNV en los nuevos casos clínicos sin caracterizar del INGEMM. Estas pruebas pueden permitir valorar si la metodología aquí desarrollada podría ser útil asistiendo a los clínicos y genetistas, ayudando a reducir los tiempos de los análisis genéticos diferenciales. Para conseguir esto, se usaron las redes de deleciones y duplicaciones antes descritas, dependiendo de la naturaleza del reordenamiento genómico en cada caso.

1) Casos clínicos aislados: El primer conjunto de casos corresponde a una cohorte de 293 pacientes (datos no publicados a la fecha de escritura de este trabajo) que presentan 519 aberraciones genéticas (312 deleciones, 155 duplicaciones y 52 reordenamientos complejos), los cuales fueron analizados haciendo uso de aCGH de oligonucleótidos o SNP *array* entre 2010 y 2014 en el IN-

GEMM. Estos pacientes fueron mayoritariamente referidos a los clínicos debido a: discapacidad intelectual, malformaciones congénitas y trastornos del espectro autista.

2) Grupo de pacientes que comparte fenotipo y genotipo, definiendo un nuevo síndrome de microdeleción/microduplicación: El segundo grupo de casos clínicos utilizado como prueba de concepto en este trabajo se obtuvo de un estudio específico de caracterización sindrómica, llevado a cabo por Nevado *et al.* [55], que incluía 13 pacientes no relacionados entre sí (con un total de 15 reordenamientos genómicos, distribuidos en 13 deleciones y 2 duplicaciones). 11 de los pacientes presentaban deleciones y 2 de ellos una única duplicación. El análisis realizado mediante aCGH junto con los registros clínicos mostraron que estos pacientes compartían características genotípicas y fenotípicas, representando un nuevo síndrome de microdeleción/microduplicación intersticial [55]. Las características comunes asociadas a este síndrome consisten en: circunferencia anormal de la cabeza (macrocefalia en los casos de deleciones y microcefalia en los de duplicaciones), discapacidad intelectual, retraso en el desarrollo, hipotonía, retraso en el habla y algunas características dismórficas.

Análisis mediante *microarrays*

La técnica *Array-CGH* se realizó utilizando una matriz de oligonucleótidos personalizados (*KaryoArray®v3.0, 8x60K, Agilent-based Technologies, Santa Clara, CA*) [187]. Brevemente, este *array* posee una densidad media de 1 sonda por cada 9Kb en las regiones clínicamente relevantes (síndromes de microdeleciones/microduplicaciones, regiones subteloméricas y pericentroméricas) y de 1 sonda por cada 175Kb en otras regiones genómicas (columna vertebral o *backbone*).

En algunos casos, se realizó un escaneo genómico de alta densidad, de 850.000 marcadores de SNP *array* en probandos, haciendo uso del diseño comercial *Illumina CytoSNP-850k BeadChip*, de acuerdo a las especificaciones del fabricante (Illumina, San Diego, CA).

3.- Generación del modelo de red

Los archivos con los datos acerca de las CNVs *de novo* de los pacientes y sus anotaciones de términos HPO fueron descargados directamente del servidor *ftp* de DECIPHER. Dado que HPO se organiza como una estructura de árbol jerárquico (figura 6.1, sección 1), cada paciente es asociado, en el modelo de red, a sus términos HPO específicos (nodos hijos) y a todos los términos parentales sobre ellos en el árbol de HPO (figura 6.1, sección 3). Un *locus* se define como una SOR (*small overlapping region*, o pequeña región solapante): la región del genoma que interseccionа con un conjunto de CNVs de pacientes de la red (figura 6.1, sección 2). Se genera un modelo de red pacientes-*loci* (figura 6.1, sección 4) conectando pacientes específicos a *loci*. La sección 5 de la figura 6.1 muestra la integración de los fenotipos HPO (círculos), los *loci* (rectángulos) y los pacientes. De este modo, las capas de los términos HPO y de las CNVs quedan conectadas a través de la capa intermedia de pacientes, constituyendo una red *tripartita*, la cual es posteriormente dividida en 2 partes; una conteniendo las deleciones y la otra las duplicaciones. La subred de deleciones consta de 45.361 relaciones únicas entre términos HPO y pacientes y 30.038 asociaciones *locus*-paciente. Por su lado, la subred de duplicaciones presenta 17.010 conexiones únicas entre términos HPO y pacientes y 10.888 relaciones *locus*-paciente.

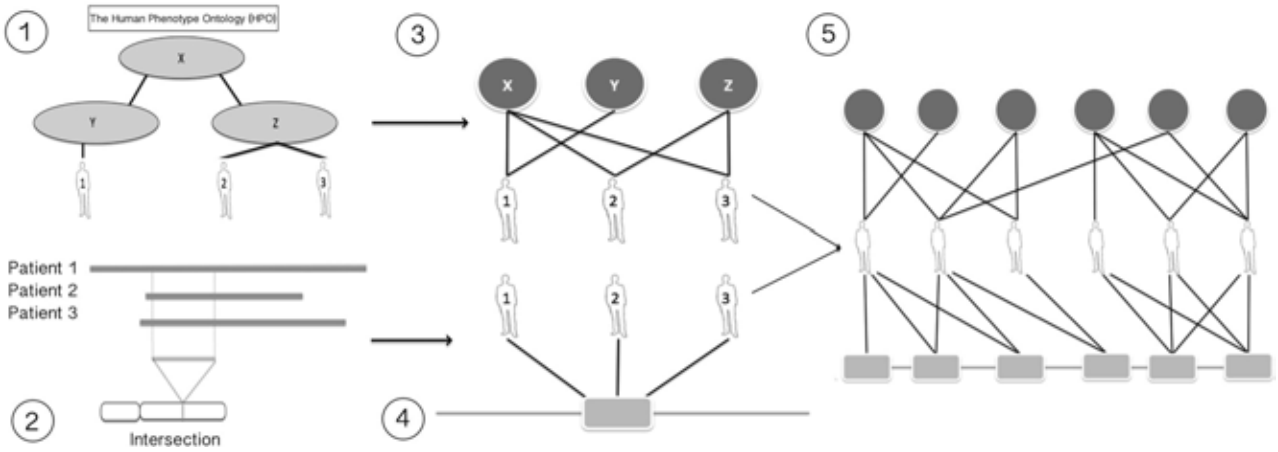


Figura 6.1: **Generación de una red *tripartita* usando los datos de pacientes de DECIPHER.** Los círculos representan fenotipos y los rectángulos *loci*. 1) Los pacientes son etiquetados fenotípicamente haciendo uso de términos HPO; 2) Se define un *locus* como la región cromosómica en la cual se solapan un conjunto de CNVs de pacientes; 3) Las capas HPOs-pacientes; 4) Las capas pacientes-*loci*; 5) La red *tripartita* final.

4.- Cálculo de la medida de asociación fenotipo-genotipo

Con el fin de medir el grado de asociación entre los términos HPO y los *loci* a través de los nodos de los pacientes en las redes *tripartitas*, se aplicó el Índice Hipergeométrico (*Hypergeometric Index* o HyI) (véase la figura 6.2). En este trabajo, el HyI proporciona la probabilidad (transformada mediante el logaritmo negativo $[-log(p)]$) de obtener un grado de asociación igual o superior al esperado por azar entre un fenotipo-*locus* dado [67]. Este índice es usado frecuentemente para medir la significancia estadística de las asociaciones en diferentes áreas: análisis funcional de *microarrays*, comparaciones computerizadas entre pares de imágenes, vectores de datos y análisis espacial en espectrometría de masas [188–190]. En la sección 1 del material suplementario, capítulo 6.3.5, se explica en detalle el método matemático, el comportamiento del algoritmo, se compara con otras métricas, y también se incluyen las instrucciones para acceder al código fuente desarrollado. Adicionalmente se llevó a cabo una validación cruzada del método (sección 2 del material suplementario, capítulo 6.3.5), se comprobaron sus posibles dependencias (sección 3 del material suplementario, capítulo 6.3.5), encontrando: **a)** una relación negativa entre los valores HyI y la fre-

cuencia de los términos HPO; **b)** una correlación positiva entre la prevalencia HPO y el número de *loci* asociados; y **c)** una ausencia de correlación entre los valores HyI y los pacientes/CNVs por *locus*. Y finalmente, se muestra un ejemplo, analizando un fenotipo prevalente (sección 4 del material suplementario, capítulo 6.3.5).

$$HyI_{AB} = -log_{10} \sum_{i=|N(A)\cap N(B)|}^{\min(|N(A),N(B)|)} \frac{\binom{|N(A)|}{i}\cdot\binom{n_y-|N(A)|}{|N(B)|-i}}{\binom{n_y}{|N(B)|}}$$

Figura 6.2: **Ecuación del Índice Hipergeométrico (HyI).** A representa un nodo de fenotipo y B un nodo de *locus* en la red tripartita. n_y es el número total de nodos en la red intermedia (pacientes) y $|N(X)|$ hace referencia al grado del nodo X (el número de nodos con el que interactúa). Para más información se puede consultar la sección 1 del material suplementario, capítulo 6.3.5 o la sección de Materiales y métodos generales donde se detallan las métricas de análisis de redes heterogéneas, capítulo 3.2.2.

El nivel de significancia estadística de la asociación se incrementa con el valor HyI, ya que a una menor probabilidad de que la asociación del fenotipo-*locus* observado sea debida al azar se obtiene un mayor valor de HyI. Esta métrica también amortigua los efectos de las CNVs largas que se solapan con muchas CNVs pequeñas, y lo hace de 2 formas: **1)** un fenotipo ampliamente distribuido en la red lleva a valores de HyI bajos y **2)** el fenotipo será calificado con altos valores únicamente si es compartido por un gran número de pacientes en el mismo *locus* (SOR). Ambos hechos, relacionados con la especificidad, son discutidos en las secciones 3 y 4 del material suplementario (capítulo 6.3.5). Un ejemplo del funcionamiento del HyI se muestra en la figura 6.3, con 2 posibles escenarios. El escenario 1 en la figura 6.3 muestra 1 fenotipo conectado a 1 *locus* a través de 3 pacientes diferentes, pero dicho fenotipo tiene 4 conexiones más a otros pacientes con trastornos genéticos localizados en *loci* diferentes (fenotipo HPO altamente prevalente). El valor de asociación HyI obtenido en este caso es bajo (HyI = 0,001; p-valor = 0,99). En el escenario 2 de la figura 6.3, 1 fenotipo está también conectado a 1 *locus* mediante 3 pacientes, pero en este caso la

cantidad de conexiones a otros pacientes que apuntan a *loci* distintos para este fenotipo es baja; este último caso representa una asociación fenotipo-*locus* más específica, y por lo tanto la significancia es mayor (HyI = 0,942; p-valor = 0,11). Con la finalidad de establecer un umbral de significancia, para este estudio, se considerarán los valores de HyI $\geq$ 2,0 como asociaciones significativas fenotipo-*locus* (un p-valor $\leq$ 0,01 de ser debido al azar).

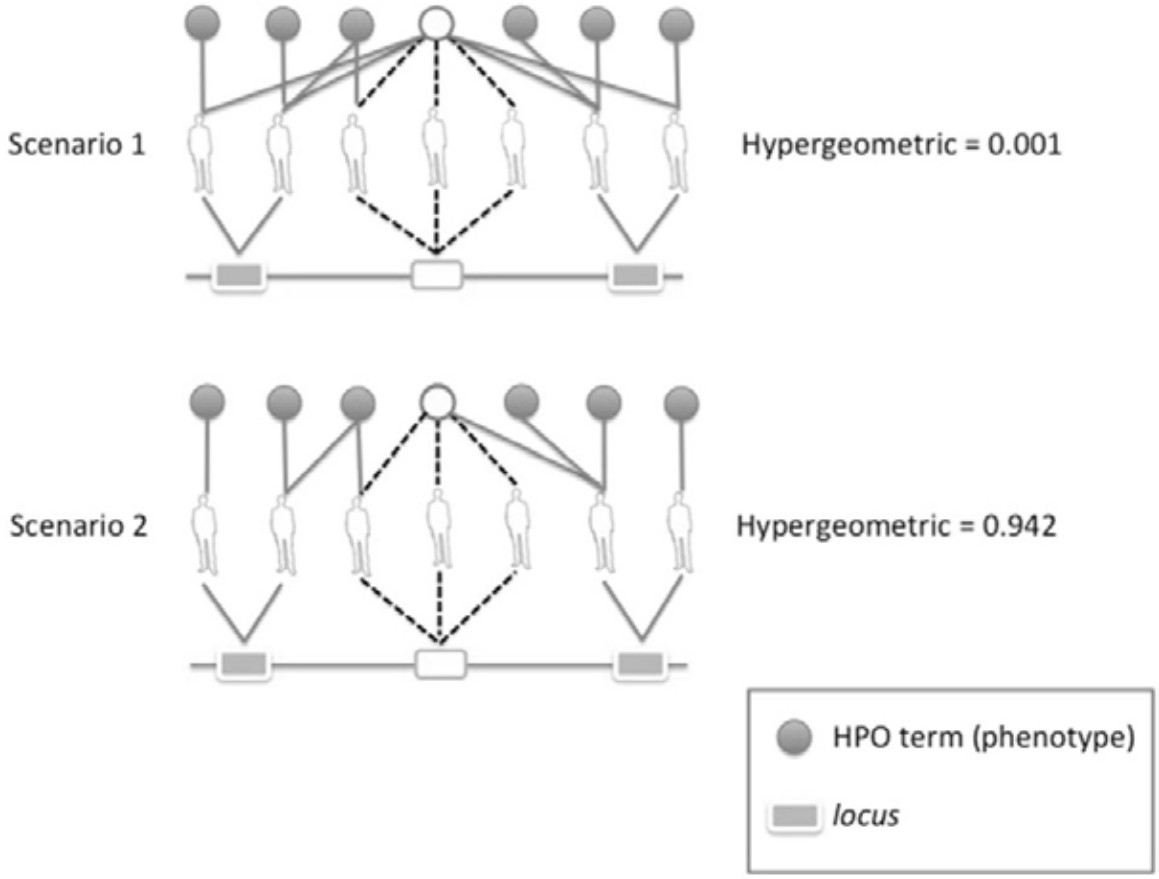


Figura 6.3: **Cálculo del Índice Hipergeométrico (HyI) en 2 escenarios en una red *tripartita*.** Los escenarios 1 y 2 muestran los valores HyI para fenotipos de alta y baja prevalencia, respectivamente; conectados a 1 *locus* a través de 3 pacientes.

5.- Midiendo el índice de asociación entre fenotipos y *loci* en la red de DECIPHER

Para calcular la significancia estadística de las asociaciones entre los fenotipos HPO y los *loci*, se aplicó el Índice Hipergeométrico (HyI) a las subredes de deleciones *de novo* y duplicaciones *de novo* de la red *tripartita* (véase la sección anterior). Se calculó el valor de asociación HyI para 600.234 pares diferentes de términos HPO-*loci* haciendo uso de la subred de deleciones y 175.956 pares mediante la de duplicaciones. En la tabla 6.1 se muestran algunos ejemplos de pares fenotipo HPO-*locus* con valores altos de asociación en la red.

HPO code	Phenotype	Max HyI (del.)	Max HyI (dup.)	*Locus* coordinates (del.)	*Locus* coordinates (dup.)
HP:0002813	Abnormality of limb bone morphology	4.93	3.10	Ch 7: 41613503–42807486	Ch 9: 11818351–12709928
HP:0000284	Abnormality of ocular region	3.39	1.65	Ch 7: 41518389–41613502	Ch Y: 945080–2654860
HP:0000153	Abnormality of the mouth	3.50	1.94	Ch 2: 200208169–200246437	Ch 22: 40849826–41082043
HP:0001315	Reduced tendon reflexes	3.16	3.54	Ch 3: 6036656–6045520	Ch 20: 29462074–29833608
HP:0010477	Aplasia of the bladder	–	3.36	–	Ch 17: 34817222–34817420
HP:0001933	Subcutaneous hemorrhage	3.38	–	Ch 21: 15398168–15412670	–
HP:0200008	Intestinal polyposis	3.56	–	Ch 10: 89717525–93614902	–
HP:0001789	Hydrops fetalis	3.56	–	Ch 13: 80378611–80386671	–
HP:0003186	Inverted nipples	2.90	3.54	Ch X: 455566–544731	Ch 16: 75683739–78186860
HP:0000699	Diastema	3.08	3.36	Ch 5: 13750113–14064732	Ch 7: 2290686–2996437
HP:0010761	Broad columella	3.56	–	Ch 19: 48066340–48270667	–
HP:0008110	Equinovarus deformity	3.26	3.36	Ch 16: 2038810–2124458	Ch 16: 90148342–90148393
HP:0002323	Anencephaly	–	3.36	–	Ch 17: 34817222–34817420

Tabla 6.1: **Ejemplos de asociaciones fenotipo HPO *vs.* *locus* identificadas en las redes de DECIPHER con valores altos.** Columnas: (1) Código HPO; (2) Descripción del fenotipo; (3) Máximo valor de HyI obtenido para el fenotipo, asociado a un *locus* en la subred de deleciones *de novo*; (4) Máximo valor de HyI obtenido para el fenotipo, asociado a un *locus* en la subred de duplicaciones *de novo*; (5) Identificador del cromosoma y coordenadas de inicio y fin (en *bps* en el genoma de referencia hg19) del *locus* asociado con el fenotipo con el máximo valor de HyI en la subred de deleciones *de novo*; y (6) en la subred de duplicaciones *de novo*.

6.- Prelación de potenciales asociaciones fenotipo/CNV en nuevos casos clínicos

Se calcularon los valores de HyI para las subredes de deleciones y duplicaciones *de novo*, y a partir de los resultados, se construyó una base de datos incluyendo todos los valores de HyI para todos los fenotipos HPO en combinación con todos los *loci*. Posteriormente, se analizó este conjunto de datos para identificar asociaciones fenotipo-genotipo, ordenadas por su valor de HyI, para nuevos pacientes que no estaban incluidos en el sistema (véase la figura 6.4).

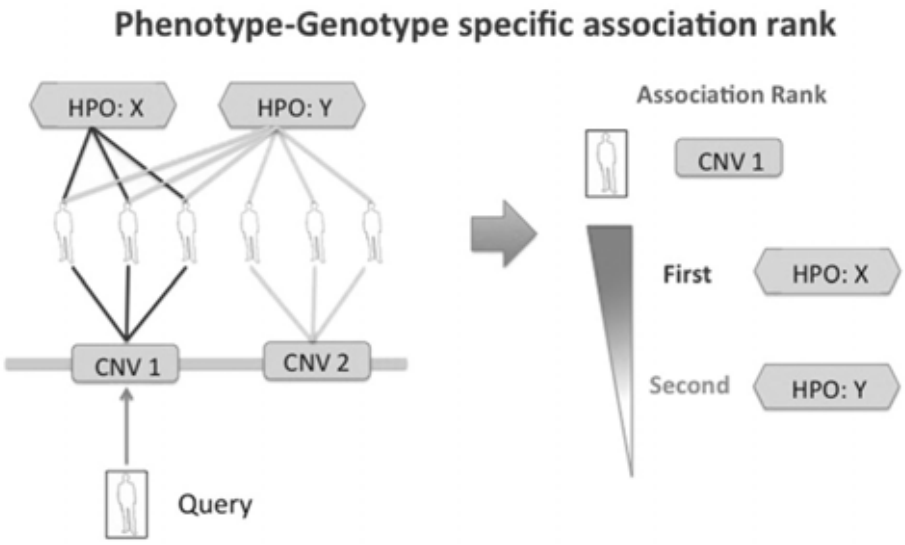


Figura 6.4: **Identificación de asociaciones fenotipo-*locus* en nuevos casos clínicos.** Una CNV de un nuevo paciente (*query*) es asignada a un *locus* (CNV 1) en la red tripartita mediante comparación de solapamientos genómicos (parte izquierda de la figura). Todos los fenotipos asociados a pacientes en ese *locus* son ordenados en base a su valor de asociación HyI en la red DECIPHER (parte derecha).

7.- Otros parámetros utilizados

Adicionalmente al valor de asociación HyI, también se definieron y calcularon algunos parámetros suplementarios con la finalidad de estimar con mayor precisión la significancia de los resultados:

Penetrancia: se define como el porcentaje de pacientes en la base de datos DECIPHER con el mismo *locus* afectado que comparten el mismo término HPO. Una ***penetrancia*** del 100 % significa que todos los pacientes con el *locus* afectado presentan el fenotipo. La penetrancia es un parámetro útil para medir los efectos de aberraciones genéticas combinadas y la probabilidad de presentar ese fenotipo asociado si un paciente sufre una mutación en esa región específica.

% max: El ratio entre el valor de HyI obtenido para un término HPO asociado a un *locus* y el máximo valor de HyI en la red para ese fenotipo. Este parámetro estima la significancia relativa de una puntuación de HyI a un fenotipo HPO dado en relación a un *locus* particular comparado con el máximo valor observado para el fenotipo en todo el genoma. Un ***% max*** del 100 % significa que el valor del HyI entre ese fenotipo HPO y ese *locus* es el máximo encontrado para ese término HPO en el genoma entero, con los datos de DECIPHER.

Solapamiento: El porcentaje de ***solapamiento*** (en pares de bases) que una CNV de un nuevo caso clínico tiene con un *locus* presente en la red de referencia. Si este parámetro es del 100 %, eso indica que la CNV estudiada está contenida en la región del *locus* (en la SOR). Este parámetro debe ser tenido en cuenta en la interpretación de los datos, ya que si el paciente analizado solamente comparte un pequeño porcentaje de su zona afectada con el *locus* (SOR) de la red, puede que no esté afectada la región específica responsable de un fenotipo particular asociado a ese *locus* (incluso siendo alto el valor de HyI entre el *locus* y el fenotipo).

6.3. Resultados y Discusión

6.3.1. Aplicación de las asociaciones encontradas en el análisis de las redes *tripartitas* a nuevos casos clínicos

Se hizo uso de 293 casos clínicos (pacientes), asociados a 519 CNVs, diagnosticados por el INGEMM (Instituto de Genética Médica y Molecular, Hospital La Paz, Madrid, España) durante el periodo comprendido entre 2010 y 2014. Para las CNVs asociadas a 258 de los 293 casos clínicos (88 %), el procedimiento encontró solapamiento con al menos un *locus* patológico y devolvió, para cada uno, una lista de fenotipos HPO asociados y ordenados por su valor de HyI (tal como se ha descrito en la figura 6.4 del capítulo 6.2 -Material y métodos-). Solo aquellos términos HPO asociados a un valor de HyI $\geq$ 2,0 (p-valor $\leq$ 0,01) fueron tenidos en cuenta. 17.096 asociaciones significativas (HyI $\geq$ 2,0) fueron encontradas, implicando a 856 fenotipos diferentes. Un total de 381 de los 1.489 (26 %) términos HPO diagnosticados por los clínicos fueron también identificados por el sistema asociado a las CNVs de los pacientes en la subred de deleciones *de novo*, y 252 de los 609 (41 %) en la subred de duplicaciones *de novo* (véase la tabla 6.2). Por otro lado, un total de 521 y 376 fenotipos HPO no diagnosticados fueron identificados por el sistema asociados, respectivamente, a CNVs en deleciones y duplicaciones *de novo* en los casos clínicos (véase la tabla 6.2). Estos resultados indican que esta nueva aproximación podría ser aplicada extensivamente en diagnósticos diferenciales a nuevos casos clínicos, con la finalidad de encontrar fenotipos asociados a CNVs concretas a través de la comparación con la información completa de los pacientes integrados en la red generada a partir de DECIPHER.

# HPO phenotypes diagnosed by the INGEMM clinicians for all the patients	1694
# Diagnosed HPOs for all the patients presenting a deletion	1489
# Diagnosed HPOs for all the patients presenting a duplication	609
# Diagnosed HPOs also identified by the system using the de novo deletions network	381
# Diagnosed HPOs also identified by the system using the de novo duplication network	252
# HPOs identified by the system and not diagnosed (with HyI > 2, penetrance 100%, and loci overlap 100%) using the de novo deletions network	521
# HPOs identified by the system and not diagnosed (with HyI > 2, penetrance 100%, and loci overlap 100%) using the de novo duplications network	376

Tabla 6.2: **Estadísticas de la comparación entre los registros clínicos de los 293 pacientes con trastornos genéticos raros de CNVs y las asociaciones fenotipos HPO-*loci* identificadas por el sistema.**

Con el fin de ilustrar la utilidad potencial de esta metodología para asistir en el diagnóstico genético, se seleccionaron y analizaron en detalle 3 casos clínicos (2 deleciones y 1 duplicación) de la cohorte de pacientes (véase la tabla 6.3). Cabe destacar que los pacientes presentados como ejemplo en la tabla 6.3 muestran un alto nivel de coincidencias entre los fenotipos diagnosticados por los clínicos y aquellos identificados mediante el sistema de asociación en red aquí expuesto. No obstante, en el primero de los casos el sistema encuentra 2 fenotipos que no fueron reportados por los clínicos: 'Macrocefalia' y 'Anormalidad de la movilidad articular' (marcados como *not reported* en la tabla 6.3). Aunque la *penetrancia* y el parámetro *%max* son relativamente bajos para estos fenotipos, los resultados sugieren llevar a cabo tests clínicos adicionales para confirmar o descartar los citados fenotipos en el paciente.

Patient 1	Deletion	chr 12	Mutation start hg19: 30 946 782		Mutation end hg19: 132 246 215
Observed phenotypes	**Associated phenotypes**	**HyI rank**	**Penetrance**	**% max**	**Node overlap**
Abnormal facial shape (HP:0001999)	+ Low-set ears (HP:0000369)	5.41	100%	100%	100%
	-> Abnormal location of ears (HP:0000357)	5.16	100%	100%	100%
	-> Abnormality of the outer ear (HP:0000356)	4.46	100%	100%	100%
	-> Abnormality of the ear (HP:0000598)	4.01	100%	100%	100%
Not reported	*+ Macrocephaly (HP:0000256)*	*3.63*	*75%*	*100%*	*100%*
	+ Abnormality of joint mobility (HP:0011729)	*3.34*	*75%*	*85%*	*100%*
	-> Abnormal joint morphology (HP:0001367)	*3.09*	*75%*	*63%*	*100%*
Global developmental delay (HP:0001263)	+ Abnormality of body weight (HP:0004323)	3.28	100%	100%	100%
	-> Growth abnormality (HP:0001507)	2.68	100%	67%	100%
	+ Short stature (HP:0004322)	3.03	100%	70%	100%
	-> Abnormality of body height (HP:0000002)	2.70	100%	72%	100%
	-> Growth delay (HP:0001510)	2.63	100%	50%	100%

Patient 2	Deletion	chr 17	Mutation start hg19: 34 911 952		Mutation end hg19: 36 510 799
Observed phenotypes	**Associated phenotypes**	**HyI rank**	**Penetrance**	**% max**	**Node overlap**
Multicystic kidney dysplasia (HP:0000003)	+ Abnormality of the kidney (HP:0000077)	3.18	50%	71%	100%
	-> Abnormality of the genitourinary system (HP:0000119)	2.87	67%	72%	100%
	-> Abnormality of the upper urinary tract (HP:0010935)	2.94	50%	70%	100%
	-> Abnormality of the urinary system (HP:0000079)	2.45	50%	66%	100%
Fetal choroid plexus cysts (HP:0011426)	*Not found*	*Not found*	*Not found*	*Not found*	*Not found*

Patient 3	Duplication	chr 16	Mutation start hg19: 22 369 809		Mutation end hg19: 22 436 522
Observed phenotypes	**Associated phenotypes**	**HyI rank**	**Penetrance**	**% max**	**Node overlap**
Abnormal facial shape (HP:0001999)	+ Depressed nasal bridge (HP:0005280)	3.45	100%	100%	100%
	+ Deviated nasal septum (HP:0004411)	3.04	33%	100%	100%
	-> Abnormality of the nasal bridge (HP:0000422)	2.48	100%	100%	100%
	-> Abnormality of the nose (HP:0000366)	2.21	75%	80%	100%
	-> Abnormality of the midface (HP:0000309)	2.37	50%	89%	100%
	+ Malar flattening (HP:0000272)	3.14	50%	91%	100%
	-> Abnormality of the zygomatic arch (HP:0005557)	3.13	50%	91%	100%
	Brittle hair (HP:0002299)	2.82	50%	100%	28%
Global developmental delay (HP:0001263)	*Not found*	*Not found*	*Not found*	*Not found*	*Not found*

Tabla 6.3: **Ejemplo del análisis de CNVs para 3 pacientes del INGEMM.** Cabecera: Coordenadas cromosómicas de la CNV (mapeadas en el genoma de referencia hg19). Columnas, de izquierda a derecha: **(1) Fenotipos observados** en el paciente descritos por el médico; **(2) Fenotipos asociados**: la lista de fenotipos identificados en la red con un nivel de asociación significativo (valor de HyI) con el *locus* que presenta solapamiento con la CNV del paciente. '+' indica que el término HPO es el más específico de entre el grupo (término hijo). '>' se utiliza para los términos parentales detectados que están relacionados con el término hijo (términos menos específicos) en la ontología HPO; **(3) Clasificación HyI (Índice Hipergeométrico); (4) Penetrancia; (5) % max; (6) Solapamiento.**

La CNV del paciente 2 tiene una región de alrededor de 100 Kb que no coincide con ningún *locus* de la red (coordenadas de la región no coincidente -hg19-: 36,410,558-36,510,799 bps en chr17). En este caso, el paciente 2 fue diagnosticado de 'Quistes de plexo coroideo fetal', una anomalía consistente en pequeñas estructuras llenas de fluido en el coroides de los ventrículos laterales del cerebro fetal. Estos resultados sugieren que la región de la CNV de este caso clínico que no coincide con ninguna de las del sistema podría contener la causa genética de los 'Quistes de plexo coroideo fetal' diagnosticados en este paciente. Esto muestra un potencial uso de la aproximación sistémica en red para discriminar las regiones particulares asociadas con los diferentes fenotipos.

Finalmente, se muestra un ejemplo de un paciente que presenta 1 duplicación (paciente 3 en la tabla 6.3). En este paciente, los clínicos observaron 2 fenotipos: 'Anomalía en la morfología facial' y 'Retraso global del desarrollo'. El segundo de ellos no fue detectado por el sistema, pero para el primero de ellos se obtienen hasta 8 fenotipos ontológicamente relacionados con un valor de asociación de HyI significativo. Los más específicos fueron los siguientes: 'Puente nasal deprimido', 'Septum nasal desviado', 'Aplanamiento malar' y 'Cabello quebradizo'. Cabe destacar que, aunque 'Cabello quebradizo' tiene un HyI significativo (2,82) muestra valores bajos de *penetrancia* y *solapamiento*, de manera que debe ser considerado con cautela. Teniendo en cuenta el resto de la información, se puede inferir que la 'Anomalía en la morfología facial' observada en este paciente podría tener relación con malformaciones de la nariz y con la estructura del arco cigomático.

6.3.2. Aplicación de la metodología a un grupo de pacientes que comparten un nuevo síndrome no recurrente de microdeleción/microduplicación

La misma aproximación descrita en la sección anterior para los casos clínicos aislados fue aplicada también a un conjunto de 13 pacientes con 15 CNVs (13 deleciones y 2 duplicaciones). Estos pacientes fueron previamente clasificados, en la unidad clínica del INGEMM, dentro de un nuevo síndrome de microdeleción/microduplicación localizado en la región genómica 19p13.3 [55]. Todos estos pacientes comparten un conjunto de fenotipos relacionados con un reordenamiento de

CNV en una misma región cromosómica (deleciones y duplicaciones). El resumen de la comparación entre los fenotipos establecidos por Nevado *et al.* [55] y los resultados del análisis sistémico se muestra en la tabla 6.4. Estos resultados revelan que la aproximación en red fue capaz de identificar 37 de las 178 asociaciones fenotipo-paciente diagnosticadas para este síndrome (21 %) con valores HyI significativos (HyI $\geq$ 2,0; p-valor $\leq$ 0,01). A pesar de la recomendación de utilizar valores de HyI superiores a 2,0 para obtener resultados altamente fiables, el sistema también proporciona, como información adicional, resultados con valores más bajos, los cuales han de ser tenidos en cuenta con precaución y evaluados para cada caso en particular, considerando información adicional antes de hacer cualquier inferencia. En este sentido, 91 de las 178 asociaciones fenotipo-paciente diagnosticadas para este síndrome (51 %) fueron también detectadas por el sistema pero con valores de HyI $<$ 2,0. Algunos de estos casos corresponden a fenotipos prevalentes, tales como 'Retraso en el desarrollo psicomotor' o 'Discapacidad intelectual'. Se diferencian ambos tipos de resultados haciendo uso de distintos colores en las celdas de la tabla 6.4. Existe un conjunto de fenotipos diagnosticados en la mayoría de los pacientes con este síndrome que fue también recurrentemente encontrado por la aproximación sistemática, con valores HyI significativos. Por ejemplo: 'Puente nasal ancho' (9 asociaciones encontradas por el sistema frente a 10 pacientes diagnosticados), 'Reflujo gastroesofágico' (4 de 4 diagnosticados), 'Hernias umbilicales' (4 de 4), 'Enfermedad cardíaca' (5 de 7) y 'Problemas con la alimentación' (5 de 6). Tal como se ha mostrado previamente, el sistema también encontró fenotipos asociados con esas CNVs en el 46 % de los pacientes con este síndrome que no han sido reportados en los registros clínicos de los pacientes, tales como: 'Anomalía del riñón', 'Anomalía del pene' y 'Anomalía del tejido conectivo'. En una revisión retrospectiva (llevada a cabo después de la aplicación de este método para verificar las predicciones directamente en pacientes) de 38 de esos pacientes con microdeleciones en 19p13.3, se encontraron anomalías renales en el 26,31 % de ellos, anomalías en los órganos sexuales en el 21,05 % y no hubo casos conocidos de anomalías en el tejido conectivo. Estos resultados respaldan el potencial del sistema desarrollado en este trabajo como asistente para el diagnóstico clínico (véase la sección 5 en el material suplementario, capítulo 6.3.5).

Patients	1	2	3	4	5	6	7	8	9	10	11	12	13
Type of CNV	-	-	-	-	-	-	-	+	-	+	-	-	-
Gender	F	F	M	F	M	M	F	F	F	F	M	F	F
Growth and development													
Psychom. Develop. delay (HP:0002194, HP:0011342, HP:0011344)	X	X	X	X	X	X	X	X	X	X	X	X	X
Intellectual disability (HP:0001256, HP:0001249)	X	X	X	X	X	X	X	X	X	X	X	X	X
Speech delay (HP:0000750)	X	X	X	X	X	X	X	X		X			X
Macro- /Microcephaly (HP:0000256, HP:0000252)	X	X	X	X	X		X	X	X		X	X	X
Overgrowth synd. testing (HP:0001548)		X			X								X
Proportionate short stature (HP:0003508)	X		X	X		X		X		X			
Abnormality of the face													
Hypertelorism (HP:0000316)	X	X	X	X	X	X			X	X			X
Downslanting palpebral fissures			X		X	X				X			
Ptosis (HP:0000508)						X							
Epicanthal folds (HP:0000286)	X	X								X			
Wide nasal bridge (HP:0000431)	X	X	X	X	X	X	X	X		X			X
Depressed nose and root (HP:0005280)	X	X	X			X	X			X			X
Philtrum anomalies (HP:0000322)	X	X		X	X			X					
Thin upper lip (HP:0200086)	X	X	X	X	X	X		X					
Ear anomalies (HP:0000598)	X	X	X	X	X	X				X			X
High or prom. forehead (HP:0011220)	X	X	X	X	X	X	X	X	X	X			X
Neurology													
Hypotonia (HP:0001290)	X	X	X	X	X	X	X	X		X	X	X	X
Behaviour (HP:0000708, HP:0000718)	X					X		X	X		X		
Hearing problems (HP:0000365)													X
Others													
Urinary reflux (HP:0000076)	X		X										
Gastroesophageal reflux (HP:0002020)			X		X		X						X
Abnormal fingers/toes (HP:0006101, HP:0001780, HP:0001167)	X		X	X				X		X			X
Feeding problems (HP:0008872)	X		X	X			X			X			X
Ophthalmologic abnormalities (HP:0000504)	X	X	X	X			X			X			
Umbilical hernias (HP:0001537)			X		X		X						X
Sleeping disorders (HP:0002360)					X		X						X
Heart disease (HP:0001627)			X		X	X	X		X	X			X
Found by the system but not diagnosed													
Abnormality of the kidney													
Abnormality of the penis													
Abnormality of connective tissue													

Tabla 6.4: **Resultados de la aplicación del método para los pacientes con el síndrome asociado a CNVs en la región 19p13.3.** La tabla muestra en sus 3 primeras filas: los identificadores de los pacientes; el tipo de CNV: '-' para deleciones y '+' para duplicaciones; y el sexo. Primera Columna: descripciones de fenotipos y códigos HPO. Las casillas con *X* indican que el fenotipo ha sido previamente diagnosticado en el paciente mediante el examen clínico. Los fenotipos encontrados por la aproximación sistémica con valores de HyI significativos (HyI $\geq$ 2,0; p-valor $\leq$ 0,01) son representados por casillas de color gris oscuro, y aquellos detectados con valores de HyI más bajos mediante casillas gris claro. Los fenotipos se han agrupado en 4 categorías generales: crecimiento y desarrollo, neurología, otros, y una categoría extra para aquellos encontrados por el sistema pero que no habían sido diagnosticados previamente en el examen clínico.

6.3.3. Discusión

Es sabido que la identificación de nuevos síndromes está basada en el establecimiento de relaciones precisas fenotipo-genotipo; no obstante, en el caso de algunas CNVs, la variabilidad en la expresión y la penetrancia de las manifestaciones clínicas han complicado la interpretación de su significancia clínica. Desde un punto de vista histórico, antes del uso comparativo de los reordenamientos cromosómicos coincidentes en el genoma, la identificación de los síndromes se basaba exclusivamente en una detallada y precisa descripción fenotípica de los pacientes. La introducción de tecnologías de escaneo del genoma completo, tales como aCGH, permite la identificación de nuevos desequilibrios en individuos que no presentan, aparentemente, características clínicas patognomónicas. El uso de técnicas de *microarrays* podría también facilitar la identificación de los genes responsables de los síndromes conocidos para los cuales las causas genéticas permanecen ocultas.

Actualmente, la identificación de nuevos síndromes podría comenzar con el reconocimiento de genotipos solapantes: un método 'genotipo primero', según el cual los pacientes son caracterizados por aberraciones genómicas similares antes de llevar a cabo una comparativa de la clínica común. Esta aproximación ha demostrado ser exitosa si se tiene en cuenta la creciente lista de nuevos síndromes de microdeleción/microduplicación descritos de esta manera en los últimos años [55, 56, 191–194]. El interés real de los clínicos y los genetistas es el de reducir el impacto de las entidades sindrómicas genéticas no reconocidas en un paciente. En este sentido, es sabido que muchos síndromes de microdeleción y microduplicación no recurrentes tienen como resultado algunas características comunes, no específicas, también presentes en muchos otros síndormes de microdeleción/microduplicación, complicando el diagnóstico.

En este trabajo se ha desarrollado y detallado una nueva aproximación sistémica que establece relaciones entre genotipos (haciendo uso de CNVs) y fenotipos (mediante términos HPO) con la finalidad de ayudar en el diagnóstico de síndromes genómicos raros. Se hizo uso de HPO (Human Phenotype Ontology) [52], la cual proporciona un conjunto de más de 13.000 clases (términos)

correctamente estructurados, exhaustivos, y bien definidos que sirven para clasificar las anomalías fenotípicas descritas en patologías humanas [10]. Una gran cantidad de algoritmos y de herramientas computacionales ya utilizan y explotan HPO. De hecho, es útil para diagnósticos clínicos diferenciales, así como para la priorización de genes candidatos asociados a enfermedades en estudios de secuenciación del exoma [119]. Como ejemplo, la aplicación web Phenomizer [195], la cual analiza relaciones entre anomalías fenotípicas humanas y enfermedades, hace uso de la base de datos HPO.

La aproximación sistemática y matemática implementada en este trabajo es capaz de establecer con gran precisión relaciones entre fenotipos y *loci* específicos, por medio de la explotación de redes de asociación a gran escala de fenotipos y genotipos en cientos de pacientes con enfermedades raras y patologías complejas. Los clínicos pueden asociar directamente las variantes de los pacientes y sus fenotipos cuando coocurren en el mismo *locus*, pero no pueden diferenciar fácilmente el grado de especificidad y de significancia estadística de una relación para cada fenotipo asociado a cada *locus* en particular, tal como sí hace el sistema implementado en este trabajo. Los resultados apoyan claramente el uso de esta herramienta para identificar potenciales *loci* relacionados con enfermedades genéticas dentro de bases de datos de libre acceso tales como DECIPHER, en las cuales aproximadamente la mitad de sus pacientes no están actualmente asociados a síndromes genéticos concretos. La aplicación de la metodología descrita en un conjunto de casos clínicos nuevos, usada como prueba de concepto en este trabajo, ha mostrado un alto potencial a la hora de facilitar el diagnóstico de estos nuevos casos clínicos no resueltos, de ordenar los fenotipos por especificidad de asociación a cada *locus* y de identificar nuevos fenotipos potenciales. De hecho, estos fenotipos podrían sugerir exploraciones clínicas adicionales que pueden ayudar a mejorar la precisión de los diagnósticos de los pacientes y la caracterización de nuevos síndromes raros, tal como aquí se ha demostrado. Los resultados obtenidos indican que el análisis comparativo de nuevos casos clínicos con las asociaciones variante-fenotipo identificadas en la red de pacientes diagnosticados previamente podría tener importantes aplicaciones en el diseño de *arrays* personalizados y

aproximaciones NGS para el diagnóstico de variantes genéticas, así como en la búsqueda de genes candidatos asociados a las diferentes regiones genómicas mutadas observadas en los pacientes.

La mayoría de la metodología actualmente disponible para asistir en el diagnóstico genético, tal como: PhenIX [112], Phenomantics [113], eXtasy [114], PHIVE, hiPHIVE [115, 116], Phevor [117], Phen-Gen [118] u OMIM-Explorer [119] es capaz de conectar las variantes detectadas en los pacientes con variantes patológicas conocidas o genes causantes de enfermedades, de manera jerarquizada y con soporte fenotípico (véase el capítulo 1.4.4). Algunos de estos métodos (PhenIX, PHIVE, hiPHIVE u OMIM-Explorer) podrían proporcionar, directa o indirectamente, una guía para el diagnóstico de enfermedades o la priorización de variantes o genes. Mención especial merece Phenomizer [195], el cual tiene como propósito fundamental la realización de diagnósticos diferenciales mediante la comparación de enfermedades conocidas y fenotipos de pacientes, haciendo uso del catálogo de enfermedades OMIM (Online Mendelian Inheritance in Man) [103]. No obstante, OMIM-Explorer es la única herramienta que -recientemente- incorpora sugerencias sobre fenotipos para un diagnóstico presuntivo diferencial, tal como hace el sistema presentado en este trabajo. La principal diferencia entre OMIM-Explorer y la aproximación aquí presentada, basada en redes, es que el primero de ellos utiliza las anotaciones acerca de: genes, variantes, enfermedades y fenotipos disponibles en OMIM y otras bases de datos similares, mientras que en este trabajo las asociaciones variantes-fenotipos son directamente inferidas de la red de pacientes. Además, otra importante diferencia y valor añadido de este trabajo es el hecho de que hace uso de una red *tripartita* (variantes-pacientes-fenotipos) construida con una gran cantidad de CNVs patológicas *de novo* presentes en pacientes con trastornos genéticos raros procedentes de la base de datos DECIPHER. Esta característica hace que la aproximación basada en la red de pacientes de DECIPHER, aquí presentada, sea especialmente apropiada para el diagnóstico comparativo en enfermedades raras y enfermedades huérfanas con origen genético (el 80 % de todas las enfermedades raras); un campo clínico que presenta una serie de retos muy especiales derivados de la escasa disposición de datos sobre pacientes y de una carencia de medios computacionales para su análisis y ayuda al diagnóstico mediante la combinación de la genómica clínica y el fenotipado médico [196–198].

6.3.4. Publicación: *Phenotype-loci associations in networks of patients with rare disorders: application to assist in the diagnosis of novel clinical cases. European Journal of Human Genetics. 2018.*

El trabajo de investigación descrito en este capítulo dio como fruto la publicación del artículo que aquí se adjunta y que sirve como aval de esta Tesis Doctoral.

European Journal of Human Genetics (2018) 26:1451–1461
https://doi.org/10.1038/s41431-018-0139-x

ESHG

ARTICLE

Phenotype-*loci* associations in networks of patients with rare disorders: application to assist in the diagnosis of novel clinical cases

Anibal Bueno[1] · Rocío Rodríguez-López[1,2] · Armando Reyes-Palomares[3] · Elena Rojano[1] · Manuel Corpas[4] · Julián Nevado[2,5] · Pablo Lapunzina[2,5] · Francisca Sánchez-Jiménez[1,2] · Juan A. G. Ranea[1,2]

Received: 21 June 2017 / Revised: 6 February 2018 / Accepted: 6 March 2018 / Published online: 26 June 2018

Abstract

Copy number variations (CNVs) are genomic structural variations (deletions, duplications, or translocations) that represent the 4.8–9.5% of human genome variation in healthy individuals. In some cases, CNVs can also lead to disease, being the etiology of many known rare genetic/genomic disorders. Despite the last advances in genomic sequencing and diagnosis, the pathological effects of many rare genetic variations remain unresolved, largely due to the low number of patients available for these cases, making it difficult to identify consistent patterns of genotype–phenotype relationships. We aimed to improve the identification of statistically consistent genotype–phenotype relationships by integrating all the genetic and clinical data of thousands of patients with rare genomic disorders (obtained from the DECIPHER database) into a phenotype–patient–genotype tripartite network. Then we assessed how our network approach could help in the characterization and diagnosis of novel cases in clinical genetics. The systematic approach implemented in this work is able to better define the relationships between phenotypes and specific *loci*, by exploiting large-scale association networks of phenotypes and genotypes in thousands of rare disease patients. The application of the described methodology facilitated the diagnosis of novel clinical cases, ranking phenotypes by *locus* specificity and reporting putative new clinical features that may suggest additional clinical follow-ups. In this work, the proof of concept developed over a set of novel clinical cases demonstrates that this network-based methodology might help improve the precision of patient clinical records and the characterization of rare syndromes.

Introduction

Decades of advances in genomic technologies are increasing the accuracy in the field of genetic diagnosis. It is now widely accepted that deep phenotyping [1] and genotypic characterization of patients accelerates the identification of new genetic diseases and/or different disease subtypes with prognostic or therapeutic implications, as well as improves our understanding of human genetic diseases [2–4]. However, the accurate diagnosis of many genetic disorders becomes more complicated when patients show complex phenotypic profiles [5], when several genomic syndromes share clinical features among them, or when rare genetic aberrations affect an extremely low number of patients, as in rare diseases. Hence, key challenges for clinicians include the interpretation or classification of novel/extremely rare variants and the understanding of the phenotypic consequences of these genetic variations. A "genotype first" approach, in which patients are classified by a similar

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41431-018-0139-x) contains supplementary material, which is available to authorized users.

✉ Juan A. G. Ranea
ranea@uma.es

[1] Department of Molecular Biology and Biochemistry, University of Malaga, 29071 Malaga, Spain

[2] CIBER de Enfermedades Raras, ISCIII, Madrid, Spain

[3] European Molecular Biology Laboratory (EMBL) Heidelberg, Meyerhorfstrasse, 1, 69117 Heidelberg, Germany

[4] Future Business Centre, King's Hedges Road, Cambridge CB4 2HY, UK

[5] Instituto de Genética Médica y Molecular (INGEMM), IdiPAZ, Hospital Universitario La Paz, Universidad Autónoma de Madrid, Madrid, Spain

6.3.5. Material Suplementario

A continuación se presentan los resultados adicionales del estudio, publicados como material suplementario del artículo *Phenotype-loci associations in networks of patients with rare disorders: application to assist in the diagnosis of novel clinical cases. European Journal of Human Genetics. 2018.*

Capítulo 7

Discusión general de los resultados

Vivimos en una isla rodeada por un mar de ignorancia. A medida que nuestra isla de conocimiento crece, también lo hace la orilla de nuestra ignorancia.

John Archibald Wheeler

En este trabajo se han explorado y aplicado principios de la Biología de Sistemas en varios escenarios diferentes. Se han modelado sistemas biológicos en forma de redes, tratando de hacerlo de la manera más precisa posible. Posteriormente se han explotado dichas redes mediante técnicas computacionales, confirmando la hipótesis principal de esta Tesis Doctoral; es decir, comprobando que estos procesos de modelado y análisis ayudan a la caracterización y la comprensión de los sistemas biológicos en diferentes áreas: molecular (a nivel de la funcionalidad de las proteínas o genes), de pacientes y fenotipos; así como también permiten realizar predicciones con la finalidad de optimizar los recursos experimentales de laboratorio en la identificación funcional de proteínas y dianas terapéuticas, y adicionalmente asistir en la consecución de diagnósticos clínicos más precisos.

1) La explotación de la información inherente a la estructura de las redes de interacción de proteínas por medio de algoritmos de tipo *kernel*, así como la aplicación de predictores validados mediante LOO y curvas ROC, han demostrado ser métodos efectivos para establecer, en base a significancia estadística, una lista ordenada de proteínas potencialmente implicadas en un mecanismo molecular de referencia, con sus derivadas terapéuticas en el caso de analizar procesos patológicos.

Mediante el uso de redes de interacción de proteínas y algoritmos probabilísticos de análisis de las mismas se desarrolló la metodología necesaria para estudiar los sistemas moleculares implicados en la transformación maligna de células tumorales de la línea MCF10CA1a (cáncer de mama) cuando ocurrían cambios en la rigidez de la matriz extracelular [11–13], asistiendo computacionalmente en la búsqueda de nuevas proteínas implicadas en dichos sistemas.

Los resultados de los tests de validación del sistema predictor (mediante LOO y curvas ROC) mostraron unos buenos rendimientos. Por lo tanto, a nivel estadístico, se puede afirmar que la metodología es robusta.

Una de las propiedades de dicha metodología es su carácter transversal, ya que es aplicable a cualquier sistema o proceso molecular en el que se quiera realizar una predicción funcional sobre

las proteínas que puedan estar implicadas en él. Por lo tanto, y debido a la imposibilidad de llevar a cabo los experimentos pertinentes en el contexto biológico inicial de células tumorales, se hizo una adaptación del método predictivo a la identificación de nuevas dianas anti-angiogénicas.

La angiogénesis se encuentra en el punto de mira de muchas investigaciones debido a que está relacionada con diversas patologías [147], entre las que destaca el cáncer, siendo uno de los factores cruciales para la progresión tumoral y la metástasis [148, 149]. Es por ello que la búsqueda de nuevos fármacos anti-angiogénesis es una estrategia terapéutica importante en cáncer. Actualmente los fármacos conocidos en este sentido se centran en la inhibición de miembros de la familia de los factores de crecimiento endotelial vascular (VEGFs) y sus receptores, lo cual solo ha dado frutos en un limitado número de casos. Por lo tanto, identificar nuevas dianas anti-angiogénicas distintas de aquellas centradas en el VEGF y sus receptores es un objetivo terapéutico de gran importancia.

Tras aplicar la metodología a este proceso y obtener las nuevas proteínas candidatas, de las 7 que se evaluaron, los experimentos *in vitro* de silenciamiento génico y de bloqueo mediante anticuerpos específicos mostraron claramente que la proteína SOD3 (*extracellular superoxide dismutase 3*) efectivamente estaba implicada en el proceso de angiogénesis. Más concretamente, su bloqueo reducía significativamente la migración de células endoteliales e inhibía completamente la formación de estructuras tubulares endoteliales; impidiendo la angiogénesis. Se realizaron experimentos adicionales *ex vivo* e *in vivo*, los cuales reforzaban estos hechos, apuntando a que SOD3 podría ser considerada como una nueva diana terapéutica en las patologías dependientes de la angiogénesis, tales como el cáncer. Se demuestra también la conveniencia del uso de sistemas de predicción *in silico*, como los desarrollados en este trabajo, con la finalidad de guiar en el diseño experimental, ahorrando tiempo y costes mediante la priorización de candidatos.

2) La comparación de las distancias filogenéticas entre pares de proteínas de una familia con sus distancias en el interactoma (expresadas mediante medidas estadísticas en redes de proteínas) es un mecanismo que ha demostrado ser útil para el estudio de las relaciones entre su evolución molecular y funcional, permitiendo obtener en ciertos casos detalles de

posiciones moleculares claves en el reconocimiento específico de proteínas interaccionantes, presentando en el caso de la familia RAS implicaciones en el estudio de ciertas terapias farmacológicas.

En este trabajo se ha llevado a cabo un análisis exhaustivo de la relación entre la filogenia de las proteínas RAS y su localización en la red de interacciones. A esto le siguieron análisis de secuencia y estructurales de las posiciones conservadas en los pares DIRP (proteínas alejadas filogenéticamente pero cercanas en el contexto de interacciones) en los sitios de unión de RAS con sus efectores. Los análisis de secuencia de estas proteínas divergentes pero interactuantes identificaron esas posiciones clave (especialmente conservadas entre los pares DIRP), las cuales mapeaban en las regiones de unión 3D que en Ras mediaban las interacciones con muchos de sus efectores. Estos resultados sustentan la idea de que estas posiciones conservadas determinan la especificidad en el reconocimiento de sus efectores, y por tanto qué pares DIRP aparecen cercanos en el interactoma, compartiendo sus contextos de interacción.

La destacada relación de las posiciones específicas de los DIRP con los sitios de unión en Ras sugiere que mutaciones puntuales de estas posiciones en células somáticas podrían resultar en un recableado de la red Ras, llevando a estados patológicos [166], particularmente en aquellas mutaciones que afectan al interruptor de regulación de estas proteínas quinasas. Apoyando dicha posibilidad, desde hace tiempo es conocido que el cambio de únicamente un par de residuos clave entre las proteínas parálogas Ras y Ral produce el intercambio de especificidad entre sus efectores naturales [168, 169]. Uno de estos residuos intercambiados entre Ras y Ral es el I36 (tomando la secuencia de HRas como referencia), el cual se corresponde con la posición específica de los DIRP más significativa, involucrada en el mayor número de sitios de unión de complejos Ras. Otras posiciones específicas de los DIRP coinciden con conocidas regiones de unión supresoras de tumores en Ras, sugiriendo que una investigación más a fondo de las posiciones conservadas en los DIRP podría inspirar nuevas aproximaciones anti-tumorales. La metodología que se describe en este trabajo podría ser extendida al estudio de otras familias de proteínas, haciendo uso del mismo procedimiento.

Los resultados mostrados añaden una perspectiva novedosa al modelo generalmente aceptado según el cual los genes parálogos divergen a lo largo del tiempo tanto en secuencia como en función dentro del interactoma [32–34]. En este trabajo se han identificado un número significativo de pares DIRP, divergentes en secuencia pero cercanos en el contexto funcional del interactoma humano.

Además, estos hallazgos amplían la visión actual sobre el papel putativo de los genes parálogos en el desarrollo y la adaptación de redes de señalización RAS funcionales y patológicas. Adicionalmente, se pueden sacar conclusiones importantes sobre las posiciones conservadas en los DIRP, en relación a su potencial relevancia funcional para el diseño y desarrollo de nuevos inhibidores de Ras.

3) En el ámbito clínico, la construcción de una red robusta, a tres niveles, que incluya: mutaciones genómicas, pacientes con trastornos genómicos raros y fenotipos categorizados en una ontología formal; y su análisis mediante algoritmos específicos para el estudio de redes *tripartitas* es un procedimiento que ha demostrado ser capaz de establecer relaciones significativas entre regiones mutadas y fenotipos, proporcionando una metodología útil para la mejora del diagnóstico clínico en pacientes con este tipo de patologías.

La identificación de nuevos síndromes está basada en el establecimiento de relaciones precisas fenotipo-genotipo; no obstante, en el caso de algunas CNVs, la variabilidad en la expresión y la penetrancia de los síntomas han complicado la interpretación de sus implicaciones clínicas. El mayor interés de los médicos y genetistas es el de reducir el número de las entidades sindrómicas genéticas no reconocidas en un paciente. Es en este punto donde, como ya señalaba Albert-László Barabási [3], la Medicina de Sistemas se convierte en una herramienta esencial para establecer relaciones moleculares entre fenotipos patológicos, mutaciones y enfermedades.

En este trabajo se ha desarrollado una aproximación sistémica que establece relaciones entre genotipos (haciendo uso de CNVs) y fenotipos (mediante términos HPO [52]) con la finalidad de ayudar en el diagnóstico de síndromes genómicos raros. Dicha aproximación es capaz de establecer con gran precisión relaciones entre fenotipos y *loci* específicos, por medio de la explotación de

redes de asociación a gran escala de fenotipos y genotipos en cientos de pacientes con enfermedades raras y patologías complejas.

Los clínicos pueden asociar directamente las variantes de los pacientes y sus fenotipos cuando coocurren en el mismo *locus*, pero no pueden diferenciar fácilmente el grado de especificidad y de significancia estadística de una relación para cada fenotipo asociado a cada *locus* en particular, tal como sí hace el sistema implementado en este trabajo. Los resultados apoyan claramente el uso de esta herramienta para identificar potenciales *loci* relacionados con enfermedades genéticas dentro de bases de datos de libre acceso tales como DECIPHER [62], en las cuales aproximadamente la mitad de sus pacientes no están actualmente asociados a síndromes genéticos concretos. La aplicación de la metodología descrita en un conjunto de casos clínicos nuevos, usada como prueba de concepto en este trabajo, ha mostrado un alto potencial a la hora de facilitar el diagnóstico de estos nuevos casos clínicos no resueltos, de ordenar los fenotipos por especificidad de asociación a cada *locus* y de identificar nuevos fenotipos potenciales. De hecho, estos fenotipos podrían sugerir exploraciones clínicas adicionales que pueden ayudar a mejorar la precisión de los diagnósticos de los pacientes y la caracterización de nuevos síndromes raros, tal como aquí se ha demostrado. Los resultados obtenidos indican que el análisis comparativo de nuevos casos clínicos con las asociaciones variante-fenotipo identificadas en la red de pacientes diagnosticados previamente podría tener importantes aplicaciones en el diseño de *arrays* personalizados y aproximaciones NGS para el diagnóstico de variantes genéticas, así como en la búsqueda de genes candidatos asociados a las diferentes regiones genómicas mutadas observadas en los pacientes.

Este trabajo hace uso de una red *tripartita* (variantes-pacientes-fenotipos) construida con una gran cantidad de CNVs patológicas *de novo* presentes en pacientes con trastornos genéticos raros procedentes de la base de datos DECIPHER [62]. Esta característica hace que esta aproximación sea especialmente apropiada para el diagnóstico comparativo en enfermedades raras con origen genético (el 80 % de todas las enfermedades raras); un campo clínico que presenta una serie de retos muy especiales derivados de la escasa disposición de datos sobre pacientes y de una carencia

de medios computacionales para su análisis y ayuda al diagnóstico mediante la combinación de la genómica clínica y el fenotipado médico [196–198].

Capítulo 8

Conclusiones

Me parece que lo que se necesita es un equilibrio exquisito entre dos necesidades conflictivas: el mayor escrutinio escéptico de todas las hipótesis que se nos presentan, y al mismo tiempo una actitud muy abierta a las nuevas ideas.

Carl Sagan

1) El desarrollo de predictores basados en la explotación de redes de interacción de proteínas mediante algoritmos de tipo *kernel*, validados a través de LOO y curvas ROC, es un método eficaz para priorizar proteínas potencialmente implicadas en un sistema o mecanismo molecular de referencia con evidente impacto positivo en la optimización de recursos experimentales, y con potenciales implicaciones terapéuticas cuando dichos mecanismos o sistemas se encuentran asociados a procesos patológicos.

2) La comparativa de distancias filogenéticas, entre pares de proteínas de la familia de parálogos RAS, y sus distancias en el interactoma humano (expresadas mediante redes de proteínas) es un mecanismo válido para el estudio de la relación entre su evolución molecular y su contexto de interacciones, pudiendo obtener en determinados casos detalles de los cambios moleculares/estructurales que determinan a su vez diferencias o similitudes en el contexto de las interacciones que dichas proteínas RAS mantienen con terceras proteínas.

3) La construcción de una red robusta, a 3 niveles (*tripartita*), que incluye: mutaciones genómicas (CNVs), pacientes con trastornos genómicos raros y fenotipos categorizados en una ontología formal (HPO); y su análisis mediante algoritmos específicos de estudio de asociaciones en red es un procedimiento capaz de identificar relaciones significativas entre regiones mutadas y fenotipos, proporcionando una herramienta útil para la asistencia en el diagnóstico clínico de pacientes con dicho tipo de patologías.

La conclusión general de esta Tesis Doctoral es que el modelado de sistemas biológicos en forma de redes de asociación ayuda significativamente a la caracterización y comprensión de dichos sistemas, y la explotación matemática de estos modelos es útil para la elaboración de predicciones, permitiendo optimizar en costes y rendimiento las aproximaciones experimentales en laboratorio para la identificación de nuevas proteínas funcionales. Del mismo modo, los análisis de redes de asociación biomédicas, en base a datos de pacientes, permiten la construcción de herramientas capaces de asistir los diagnósticos clínicos, facilitándolos.

Capítulo 9

Conclusions

Science knows no country, because knowledge belongs to humanity, and is the torch which illuminates the world.

Louis Pasteur

1) The development of predictors based on the exploitation of protein interaction networks using *kernel* analysis algorithms, validated by LOO and ROC curves, is an effective method to prioritize proteins potentially involved in a molecular mechanism of reference, with possible therapeutic implications when pathological processes are studied.

2) The comparison of phylogenetic distances, between protein pairs of the RAS family of paralogs, and their distances in the interactome (expressed by means of protein networks) is a valid methodology to study the relationships between their molecular and functional evolution, being able to obtain, in certain cases, details of the molecular/structural changes that in turn determine differences or similarities in the context of interactions of RAS paralogs with third proteins.

3) The construction of a robust network, with 3 levels (*tripartite*), which includes: genomic mutations (CNVs), patients with rare genomic disorders and phenotypes categorized in a formal ontology (HPO); and its analysis by means of specific algorithms to study network associations is a procedure capable of identifying significant relationships between mutated regions and phenotypes, providing a useful tool for assisting the clinical diagnosis of patients with such pathologies.

The general conclusion of this Doctoral Thesis is that the modeling of biological systems in the form of association networks allows advances in the characterization of such systems, and the mathematical exploitation of these models is useful for making predictions, allowing us to optimize laboratory experimentation for the identification of new functional proteins. In the same way, analyses of biomedical association networks, based on patient data, allow the construction of tools capable of assisting and guiding clinical diagnoses.

Bibliografía

[1] Kotlyar M, Pastrello C, Pivetta F et al. *In silico prediction of physical protein interactions and characterization of interactome orphans.* Nat Methods 2014; 12: 79-84.

[2] Mitchell M. *Complex systems: Network thinking.* Artif. Intell. 2006; 170: 1194-1212.

[3] Barabási A-L, Gulbahce N, Loscalzo J. *Network medicine: a network-based approach to human disease. Nat Rev Genet 2011*; 12: 56-68.

[4] Albert R, Barabási AL. *Statistical mechanics of complex networks. Rev Mod Phys 2002*; 74, 47-97.

[5] Zhu X, Gerstein M, Snyder M. *Getting connected: analysis and principles of biological networks. Genes Dev 2007*; 21, 1010-1024.

[6] European C. *Rare diseases.* 2014.

[7] Swaminathan GJ, Bragin E, Chatzimichali EA, Corpas M, Bevan AP, Wright CF et al. *DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders. Hum Mol Genet 2012*; 21: R37-44.

[8] Nachtomy O, Shavit A, Yakhini Z. *Gene expression and the concept of the phenotype.* Stud Hist Philos Sci Part C Stud Hist Philos Biol Biomed Sci 2007; 38: 238-254.

[9] Biesecker LG. *Phenotype matters.* Nat Genet 2004; 36: 323-324.

[10] Robinson PN, Mundlos S. *The Human Phenotype Ontology. Clin Genet 2010*; 77: 525-534.

[11] Artacho-cordón A, Artacho-cordón F, Ríos-arrabal S, Calvente I, Núñez MI. *Tumor microenvironment and breast cancer progression. A complex scenario 2012.* Landes Bioscience. 2012; 13: 14-24.

[12] Levental KR, Yu H, Kass L et al. *Matrix Crosslinking Forces Tumor Progression by Enhancing Integrin Signaling.* Cell 2009; 139: 891-906.

[13] Janmey PA, Miller RT. *Mechanisms of mechanical signaling in development and disease.* 2011. doi:10.1242/jcs.071001.

[14] García-Vilas JA, Morilla I, Bueno A, Martínez-Poveda B, Medina MÁ, Ranea JAG. *In silico prediction of targets for anti-angiogenesis and their in vitro evaluation confirm the involvement of SOD3 in angiogenesis.* Oncotarget 2018; 9. doi:10.18632/oncotarget.24693.

[15] Winograd-Katz SE, Fässler R, Geiger B, Legate KR. *The integrin adhesome: from genes and proteins to human disease.* Nat Rev Mol Cell Biol 2014; 15: 273-288.

[16] Yu H, Mouw JK, Weaver VM. *Forcing form and function: biomechanical regulation of tumor evolution.* Trends Cell Biol 2011; 21: 47-56.

[17] Huang DW, Sherman BT, Lempicki RA. (2009). *Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009* - 4(1):44-57.

[18] Huang DW, Sherman BT, Lempicki RA. (2009). *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research. 2009* - 37(1):1-13.

[19] Kanehisa, M. and Goto, S. (2000). *KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 2000* - 28, 27-30.

[20] The Gene Ontology Consortium. (2000). *Gene ontology: tool for the unification of biology. Nat. Genet.. 2000* - May 2000;25(1):25-9.

[21] Razick S, Magklaras G, Donaldson IM. *iRefIndex: A consolidated protein interaction database with provenance.* BMC Bioinformatics 2008; 9. doi:10.1186/1471-2105-9-405.

[22] Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Makela, T.P., Hautaniemi, S. (2009). *Integrated network analysis platform for protein-protein interactions, Nature methods* - 6, 75-77.

[23] Jensen et al. *STRING 8–a global view on proteins and their functional interactions in 630 organisms Nucleic Acids Research. 2009* - 37(Database issue):D412-6.

[24] Croft D, Mundo A, Haw R, Milacic M. *The Reactome pathway knowledgebase.* Nucleic acids 2014; 42: D472-D477.

[25] Charlotte M. Deane and Tom L. Blundell. (2001). *CODA: A combined algorithm for predicting the structurally variable regions of protein models. The Protein Society. 2001* - March; 10(3): 599-612.

[26] Najafov J, Najafov A. *GECO: gene expression correlation analysis after genetic algorithm-driven deconvolution.* Bioinformatics 2018; : bty623-bty623.

[27] Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. *HIPPIE v2.0: Enhancing meaningfulness and reliability of protein-protein interaction networks.* Nucleic Acids Res 2017; 45: D408-D414.

[28] Lowy DR, Willumsen BM. (1993). *Function and regulation os ras. Annu Rev Biochem. 1993; 62:851-91.* - doi: 10.1146/annurev.bi.62.070193.004223.

[29] Diego Díez, Francisca Sánchez-Jiménez, Juan A.G. Ranea (2011). *Evolutionary expansion of the Ras switch regulatory module in eukaryotes. Nucleic Acids Research. 2011* - gkr154v1-gkr154.

[30] Malumbres M, Barbacid M. *RAS oncogenes: the first 30 years.* Nat Rev Cancer 2003; 3: 459-65.

[31] McCormick F. *KRAS as a Therapeutic Target.* Clin Cancer Res 2015; 21: 1797-801.

[32] Conant GC, Wolfe KH. *Turning a hobby into a job: how duplicated genes find new functions.* Nat Rev Genet 2008; 9: 938-50.

[33] Emmert-Streib F. *Limitations of gene duplication models: Evolution of modules in protein interaction networks.* PLoS One 2012; 7. doi:10.1371/journal.pone.0035531.

[34] Sun MG, Kim PM. *Evolution of biological interaction networks: from models to real data.* Genome Biol. 2011; 12: 235.

[35] J D Thompson, D G Higgins, T J Gibson. (1994). *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research. 1994* - November 11; 22(22): 4673-4680.

[36] Howe K, Bateman A, Durbin R. *QuickTree: building huge Neighbour-Joining trees of protein sequences.* Bioinformatics 2002; 18: 1546-1547.

[37] Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW. (2001). *BIND–The Biomolecular Interaction Network Database.* Nucleic Acid Research. 2001 - Jan 1;29(1):242-5

[38] Breitkreutz BJ, Stark C, Tyers M. (2003). *The GRID: the General Repository for Interaction Datasets.* Genome Biology. 2003 - 4(3):R23. Epub 2003 Feb 27.

[39] Keshava Prasad TS, Goel R, Kandasamy K et al. *Human Protein Reference Database–2009 update.* Nucleic Acids Res 2009 ; 37: D767-D772.

[40] Kerrien S, Aranda B, Breuza L et al. *The IntAct molecular interaction database in 2012.* Nucleic Acids Res 2012; 40. doi:10.1093/nar/gkr1088.

[41] Licata L, Briganti L, Peluso D et al. *MINT, the molecular interaction database: 2012 Update.* Nucleic Acids Res 2012; 40. doi:10.1093/nar/gkr930.

[42] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. *The Database of Interacting Proteins: 2004 update.* Nucleic Acids Res 2004; 32: D449-D451.

[43] Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stümpflen V, Mewes HW, Ruepp A, Frishman D. (2005). *The MIPS mammalian protein-protein interaction database. - Bioinformatics 2005.* - 21(6):832-834; [Epub 2004 Nov 5] doi:10.1093/bioinformatics/bti115.

[44] Ulrich Güldener, Martin Münsterkötter, Matthias Oesterheld, Philipp Pagel, Andreas Ruepp, Hans-Werner Mewes, Volker Stümpflen. (2006). *MPact: the MIPS protein interaction resource on yeast. - Nucleic Acids Research. 2006.* - 34(suppl 1): D436-D441 doi:10.1093/nar/gkj003.

[45] Pazos F, Ranea JAG, Juan D, Sternberg MJE. *Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome.* J Mol Biol 2005; 352: 1002-1015.

[46] Hériché J-K, Lees JG, Morilla I et al. *Integration of biological data by kernels on graph nodes allows prediction of new genes involved in mitotic chromosome condensation.* Mol Biol Cell 2014. doi:10.1091/mbc.E13-04-0221.

[47] R Development Core Team. *R: A Language and Environment for Statistical Computing. 2011.* - ISBN 3-900051-07-0.

[48] Henikoff S, Henikoff JG. *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A 1992; 89: 10915-10919.

[49] Bernstein FC, Koetzle TF, Williams GJ et al. *The Protein Data Bank. A computer-based archival file for macromolecular structures.* Eur J Biochem 1977; 80: 319-324.

[50] Kabsch W, Sander C. *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers 1983; 22: 2577-2637.

[51] Robinson PN. *Deep phenotyping for precision medicine.* Hum Mutat 2012; 33: 777-780.

[52] Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth H V, Bailleul-Forestier I et al. *The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res 2014*; 42: D966-74.

[53] Köhler S, Vasilevsky NA, Engelstad M et al. *The human phenotype ontology in 2017.* Nucleic Acids Res 2017; 45: D865-D876.

[54] Shaw-Smith C, Redon R, Rickman L, Rio M, Willatt L, Fiegler H et al. *Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. J Med Genet 2004*; 41, 241-248.

[55] Nevado J, Rosenfeld JA, Mena R, Palomares-Bralo M, Vallespín E, Ángeles Mori M et al. *PIAS4 is associated with macro/microcephaly in the novel interstitial 19p13.3 microdeletion/microduplication syndrome. Eur J Hum Genet 2015.* doi:10.1038/ejhg.2015.51.

[56] Tenorio, J. et al. *A new overgrowth syndrome is due to mutations in RNF125. Hum. Mutat. 2014*; 35, 1436-41.

[57] Roizen NJ, Patterson D. *Down's syndrome. Lancet 2003*; 361, 1281-1289.

[58] Ballif BC, Hornor SA, Jenkins E, Madan-Khetarpal S, Surti U, Jackson KE et al. *Discovery of a previously unrecognized microdeletion syndrome of 16p11.2-p12.2. Nat Genet 2007*; 39, 1071-1073.

[59] Shaw-Smith C, Pittman AM, Willatt L, Martin H, Rickman L, Gribble S et al. *Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. Nat Genet 2006*; 38, 1032-1037.

[60] Reyes-Palomares A, Bueno A, Rodríguez-López R et al. *Systematic identification of phenotypically enriched loci using a patient network of genomic disorders.* BMC Genomics 2016; 17: 232.

[61] Rodríguez-López R, Reyes-Palomares A, Sánchez-Jiménez F, Medina MA. *PhenUMA: a tool for integrating the biomedical relationships among genes and diseases.* BMC Bioinformatics 2014; 15: 375.

[62] Firth H V., Richards SM, Bevan a. P, Clayton S, Corpas M, Rajan D et al. *DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet 2009*; 84: 524-533.

[63] Bragin E, Chatzimichali EA, Wright CF, Hurles ME, Firth H V., Bevan AP et al. *DECIPHER: Database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation.* Nucleic Acids Res 2014; 42. doi:10.1093/nar/gkt937.

[64] Lu X-Y, Phung MT, Shaw CA et al. *Genomic imbalances in neonates with birth defects: high detection rates by using chromosomal microarray analysis.* Pediatrics 2008; 122: 1310-8.

[65] Veltman, J. A. & Brunner, H. G. *De novo mutations in human genetic disease. Nat. Rev. Genet. 2012*; 13, 565-75.

[66] Dolan, M. et al. *A novel microdeletion/microduplication syndrome of 19p13.13. Genet. Med. 2010*; 12, 503-11.

[67] Fuxman Bass JI, Diallo A, Nelson J, Soto JM, Myers CL, Walhout AJM. *Using networks to measure similarity between genes: association index selection. Nat Methods 2013*; 10: 1169-76.

[68] Descartes. *Discourse on the method.* Descartes. Key philosophical writings. 1637, pp 71-122.

[69] Angell JR. *Loeb's 'The mechanistic conception of life'*. J Anim Behav 1913; 3: 464-468.

[70] KERR JG. *Holism and Evolution.* Nature 1927; 119: 307-309.

[71] Von Bertalanffy L. *General System Theory.* Georg Braziller New York 1968; 1: 289.

[72] Polanyi M. *Life's Irreducible Structure: Live mechanisms and information in DNA are boundary conditions with a sequence of boundaries above them.* Science (80-) 1968; 160: 1308-1312.

[73] Jacob F. *The Logic of Living Systems.* 1974 doi:10.1007/BF00933730.

[74] Klose J. *Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals.* Humangenetik 1975; 26: 231-243.

[75] O'Farrell PH. *High resolution two-dimensional electrophoresis of proteins.* J Biol Chem 1975; 250: 4007-21.

[76] Taub EF, Deleo JM, Brad T. *Sequential comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated RNAs.* Dna 1983; 2: 309-327.

[77] Goffeau A, Barrell G, Bussey H et al. *Life with 6000 genes.* Science (80-) 1996; 274: 546-567.

[78] Marcotte EM, Xenarios I, van der Bliek AM, Eisenberg D. *Localizing proteins in the cell from their phylogenetic profiles.* Proc Natl Acad Sci 2000; 97: 12115-12120.

[79] Mazzocchi F. *Complexity in biology. Exceeding the limits of reductionism and determinism using complexity theory.* EMBO Rep. 2008; 9: 10-14.

[80] Westerhoff H V., Palsson BO. *The evolution of molecular biology into systems biology.* Nat. Biotechnol. 2004; 22: 1249-1252.

[81] Kitano H. *Systems biology: a brief overview.* Science 2002; 295: 1662-4.

[82] Civelek M, Lusis AJ. *Systems genetics approaches to understand complex traits.* Nat. Rev. Genet. 2014; 15: 34-48.

[83] Mardinoglu A, Nielsen J. *Systems medicine and metabolic modelling.* In: Journal of Internal Medicine. 2012, pp 142-154.

[84] Loscalzo J, Barabasi A-L. *Systems biology and the future of medicine.* Wiley Interdiscip Rev Syst Biol Med 2011; 3: 619-627.

[85] Ackerman JP, Bartos DC, Kapplinger JD, Tester DJ, Delisle BP, Ackerman MJ. *The Promise and Peril of Precision Medicine.* Mayo Clin Proc 2016; : 1-11.

[86] Jaime Huerta-Cepas, Hernán Dopazo, Joaquín Dopazo, Toni Gabaldón. (2007). *The human phylome. Genome Biology. 2007.* - 8:R109, doi:10.1186/gb-2007-8-6-r109.

[87] Jaime Huerta-Cepas, Anibal Bueno, Joaquín Dopazo, Toni Gabaldón. (2008). *PhylomeDB: a database for genome-wide collections of gene phylogenies. Nucleic Acids Research. 2008* - Jan;36(Database issue):D491-6. Epub 2007 Oct 25. PMID: 17962297.

[88] William H. Piel, Michael Donoghue, Mike Sanderson. (2002). *TreeBASE: A database of phylogenetic information. Proceedings of the 2nd International Workshop of Species. 2002* - 41-47.

[89] Li H. *TreeFam: a curated database of phylogenetic trees of animal gene families.* Nucleic Acids Res 2006; 34: D572-D580.

[90] Standard TN, Cayley A, The D, Standard N. *The Newick tree format.* http://evolution.genetics.washington.edu/phylip/newicktree.html 1857; : 5-7.

[91] Peer Bork, Lars J Jensen, Christian von Mering, Arun K Ramani, Insuk Lee, Edward M Marcotte. (2004). *Protein interaction networks from yeast to human. Science Direct. 2004* - Structural Biology, 14:292-299.

[92] Stumpf MPH, Thorne T, de Silva E et al. *Estimating the size of the human interactome.* Proc Natl Acad Sci U S A 2008; 105: 6959-64.

[93] Venkatesan K, Rual JF, Vazquez A et al. *An empirical framework for binary interactome mapping.* Nat Methods 2009; 6: 83-90.

[94] Ranea JAG, Morilla I, Lees JG, Reid AJ, Yeats C, et al. (2010). *Finding the 'Dark Matter' in Human and Yeast Protein Network Prediction and Modelling. PLoS Computational Biology* - 6(9): e1000945. doi:10.1371/journal.pcbi.1000945.

[95] J G Lees, J K Heriche, I Morilla, J A Ranea, C A Orengo. (2011). *Systematic computational prediction of protein interaction networks. Physical Biology. 2011* - 8 (2011) 035008 (13pp) - doi: 10.1088/1478-3975/8/3/035008.

[96] Moya-García AA, Ranea JAG. *Insights into polypharmacology from drug-domain associations.* Bioinformatics 2013; 29: 1934-1937.

[97] Roded Sharan, Igor Ulitsky, Ron Shamir. (2007). *Network-based prediction of protein function. Molecular Systems Biology 2007.* - 3:88 doi:10.1038/msb4100129.

[98] Daniel MG, Pawlik TM, Fader AN, Esnaola NF, Makary MA. *The Orphan Drug Act.* Am J Clin Oncol 2016; 39: 210-213.

[99] EURODIS. *Rare Diseases: understanding this Public Health Priority.* October 2005; : 1-14.

[100] EURORDIS Rare Diseases Europe. *What Is a Rare Disease?* Rare Dis Eur 2007; 14-15.

[101] Zarrei M, Macdonald JR, Merico D, Scherer SW, Mg C. *A copy number variation map of the human genome.* Nat Publ Gr 2015; 16: 172-183.

[102] Krakow D, Robertson SP, King LM, Morgan T, Sebald ET, Bertolotto C et al. *Mutations in the gene encoding filamin B disrupt vertebral segmentation, joint formation and skeletogenesis.* Nat Genet 2004; 36: 405-410.

[103] Amberger J, Bocchini CA, Scott AF, Hamosh A. *McKusick's online mendelian inheritance in man (OMIM).* Nucleic Acids Res 2009: 37 (Database issue): D793-D796.

[104] Fryns JP, de Ravel TJL. *London Dysmorphology Database, London Neurogenetics Database and Dysmorphology Photo Library on CD-ROM [Version 3] 2001.* Winter RM, Ba-

raitser M. Oxford University Press, ISBN 019851-780, pound sterling 1595. [Hum Genet 2002: 111:113].

[105] *POSSUM.* Retrieved from http://www.possum.net.au/ (Accessed 2015).

[106] *Orphanet: An online database of rare diseases and orphan drugs. Copyright, INSERM 1997.* Retrieved from http:// www.orpha.net (Accessed 2015).

[107] Hoehndorf, R., Harris, M. A., Herre, H., Rustici, G., and Gkoutos, G. V. (2012). *Semantic integration of physiology phenotypes with an application to the Cellular Phenotype Ontology.* Bioinformatics, 28(13):1783-1789.

[108] Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. *The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease.* Am J Hum Genet 2008; 83: 610-615.

[109] Butte AJ, Kohane IS. *Creation and implications of a phenome-genome network.* Nat Biotechnol 2006: 24: 55-62.

[110] Claustres M, Horaitis O, Vanevski M, Cotton RGH. *Time for a unified system of mutation description and reporting: A review of locus-specific mutation databases.* Genome Res 2002; 12: 680-688.

[111] Hancock, J. M. (2014). *Phenomics.* CRC Press.

[112] Torices, R. & Muñoz-Pajares, A. J. *PHENIX: An R Package to Estimate a Size-Controlled Phenotypic Integration Index. Appl. Plant Sci. 2015*; 3, 1400104.

[113] Masino AJ, Dechene ET, Dulik MC, Wilkens A, Spinner NB, Krantz ID, et al. *Clinical phenotype-based gene prioritization: an initial study using semantic similarity and the human phenotype ontology. BMC Bioinformatics. 2014*; 15:248.

[114] Sifrim A, Popovic D, Tranchevent L-C, Ardeshirdavani A, Sakai R, Konings P, et al. *eXtasy: variant prioritization by genomic data fusion. Nat Meth. 2013*; 10:1083-4.

[115] Robinson PN, Kohler S, Oellrich A, Sanger Mouse Genetics Project, Wang K, Mungall CJ, et al. *Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res. 2014*; 24:340-8.

[116] Haendel MA, Vasilevsky N, Brush M, Hochheiser HS, Jacobsen J, Oellrich A, et al. *Disease insights through cross-species phenotype comparisons. Mamm Genome. 2015*; 26:548-55.

[117] Singleton MV, Guthery SL, Voelkerding KV, Chen K, Kennedy B, Margraf RL, et al. *Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. Am J Hum Genet. 2014*; 94:599-610.

[118] Javed A, Agrawal S, Ng PC. *Phen-Gen: combining phenotype and genotype to analyze rare disorders. Nat Meth. 2014*; 11:935-7.

[119] James, R. A. et al. *A visual and curatorial approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics. Genome Med. 2016*; 8, 13.

[120] Hoffmann, R., Valencia, A. (2004). *A Gene Network for Navigating the Literature. Nature Genetics. 2004* - 36, 664.

[121] Mellor JC, Yanai I, Clodfelter KH, Mintseris J, DeLisi C. (2002). *Predictome: a database of putative functional links between proteins. Nucleic Acids Research. 2002* - 30:306-309.

[122] Ruepp A, Waegele B, Lechner M et al. CORUM: The comprehensive resource of mammalian protein complexes-2009. Nucleic Acids Res 2009; 38. doi:10.1093/nar/gkp914.

[123] Brown KR, Jurisica I. Online predicted human interaction database. Bioinformatics 2005; 21: 2076?2082.

[124] Dijkstra EW. *A note on two problems in connexion with graphs.* Numer Math 1959; 1: 269-271.

[125] Hart PE, Nilsson NJ, Raphael B. *A Formal Basis for the Heuristic Determination of Minimum Cost Paths.* IEEE Trans Syst Sci Cybern 1968; 4. doi:10.1109/TSSC.1968.300136.

[126] Floyd RW. *Algorithm 97: Shortest path.* Commun ACM 1962; 5: 345.

[127] Fouss F., Franoisse K., Yen L., Pirotte A., and Saerens M. (2009). *An Experimental Investigation of Graph Kernels on Collaborative Recommendation and Semisupervised Classification. Proceedings of the Eighth International Conference on Data Mining (ICDM 09). 2009*.

[128] P. Diaconis and P. Hanlon. (1992). *Eigen analysis for some examples of the Metropolis algorithm, Hypergeometric functions on domains of positivity, Jack polynomials, and applications. Contemporary Math. Prob.138. 1992* - Amer. Math. Soc., Providence, RI.

[129] P. Diaconis and L. Saloff-Coste. (1992). *Comparison theorems for random walk on finite groups. Ann. Prob. 21. 1993* - 2131-2156.

[130] Tolga Can, Orthan Çamoglu, Ambuj K. Singh (2005). *Analysis of Protein-Protein Interaction Networks Using Random Walks. BIOKDD 05.* - doi:10.1145/1134030.1134042.

[131] P. Diaconis and D. Stroock. (1991). *Geometric bounds for eigenvalues of Markov chains. Annals of Appl. Prob. 1991* - 1(1991), 36:62.

[132] P. G. Doyle and J. L. Snell. (1984). *Random walks and Electric Networks. MAA. 1984*.

[133] Chebotarev P, Shamis E. *The Matrix-Forest Theorem and Measuring Relations in Small Social Groups.* Autom Remote Control 1997; 58:10.

[134] Chebotarev P. *Spanning forests and the golden ratio.* Discret Appl Math 2008; 156: 813-821.

[135] Shawe-Taylor J, Cristianini N. *Kernel Methods for Pattern Analysis.* 2004 doi:10.2277.

[136] Kondor, R.I., Lafferty, J.D. (2002). *Diffusion kernels on graphs and other discrete input spaces. ICML 02: Proceedings of the Nineteenth International Conference on Machine Learning. 2002* - pp. 315-322.

[137] Smola, A., Kondor, R. (2003). *Kernels and regularization on graphs. Proceedings of the Annual Conference on Computational Learning Theory and Kernel Workshop, Lecture Notes in Computer Science. 2003*.

[138] Lanckriet, G., Deng, M., Cristianini, N., Jordan, M., Noble, W. (2004). *Kernel-based data fusion and its application to protein function prediction in yeast. Pac Symp Biocomput. 2004* - pp. 300-11.

[139] Sebastian Köhler,Sebastian Bauer,Denise Horn,Peter N. Robinson. (2008). *Walking the Interactome for Prioritization of Candidate Disease Genes. The Americal Journal of Human Genetics. 2008* - 82, 949-958, April 2008.

[140] Devijver PA, Kittler J. *Pattern recognition: a statistical approach.* 1982.

[141] Pao-Yang Chen, Charlotte M. Deane, Gesine Reinert (2008). *Predicting and Validating Protein Interactions Using Network Structure. Plos Computational Biology. 2008* - 4(7): e1000118. doi:10.1371/journal.pcbi.1000118.

[142] Lee I, Lehner B, Crombie C, Wong W, Fraser AG, Marcotte EM. *A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans.* Nat Genet 2008; 40: 181-188.

[143] Qi Y, Suhail Y, Lin YY, Boeke JD, Bader JS. *Finding friends and enemies in an enemies-only network: A graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions.* Genome Res 2008; 18: 1991-2004.

[144] Hu P, Janga SC, Babu M et al. *Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins.* PLoS Biol 2009; 7: 0929-0947.

[145] Rojas AM, Santamaria A, Malik R et al. *Uncovering the molecular machinery of the human spindle-an integration of wet and dry systems biology.* PLoS One 2012; 7. doi:10.1371/journal.pone.0031813.

[146] Lees JG, Hériché JK, Morilla I et al. *FUN-L: Gene prioritization for RNAi screens.* Bioinformatics 2015; 31: 2052-2053.

[147] Rodríguez-Caso L, Reyes-Palomares A, Sánchez-Jiménez F, Quesada AR, Medina MÁ. *What is known on angiogenesis-related rare diseases? A systematic review of literature.* J Cell Mol Med 2012; 16: 2872-2893.

[148] Carmeliet P. *Angiogenesis in life, disease and medicine.* Nature. 2005; 438: 932-936.

[149] R. Quesada A, Angel Medina M, Munoz-Chapuli R, Luis G. Ponce A. *Do Not Say Ever Never More: The Ins and Outs of Antiangiogenic Therapies.* Curr Pharm Des 2010; 16: 3932-3957.

[150] Forbes, S. A. et al. *COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2014*; 43, D805-811.

[151] Karnoub AE, Weinberg RA. *Ras oncogenes: split personalities.* Nat Rev Mol Cell Biol 2008; 9: 517-31.

[152] Hernández-Porras I, Fabbiano S, Schuhmacher AJ et al. *K-Ras V14I recapitulates Noonan syndrome in mice.* Proc Natl Acad Sci 2014; 111: 16395-16400.

[153] Mazhab-Jafari MT, Marshall CB, Smith MJ et al. *Oncogenic and RASopathy-associated K-RAS mutations relieve membrane-dependent occlusion of the effector-binding site.* Proc Natl Acad Sci U S A 2015; 112: 6625-30.

[154] Peschard P, McCarthy A, Leblanc-Dominguez V et al. *Genetic deletion of RALA and RALB small GTPases reveals redundant functions in development and tumorigenesis.* Curr Biol 2012; 22: 2063-8.

[155] Guin S, Theodorescu D. *The RAS-RAL axis in cancer: evidence for mutation-specific selectivity in non-small cell lung cancer.* Acta Pharmacol Sin 2015; 36: 291-7.

[156] EMBL, SIB Swiss Institute of Bioinformatics, Protein Information Resource (PIR). *UniProt.* Nucleic acids research. 2013, p 41: D43-D47.

[157] Pellegrini M, Haynor D, Johnson JM. *Protein interaction networks.* Expert Rev Proteomics 2004; 1: 239-249.

[158] Mostafavi S, Morris Q. *Fast integration of heterogeneous data sources for predicting gene function with limited annotation.* Bioinformatics 2010; 26: 1759-1765.

[159] Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. *Jalview Version 2-A multiple sequence alignment editor and analysis workbench.* Bioinformatics 2009; 25: 1189-1191.

[160] Forbes SA, Bindal N, Bamford S et al. *COSMIC: Mining complete cancer genomes in the catalogue of somatic mutations in cancer.* Nucleic Acids Res 2011; 39. doi:10.1093/nar/gkq929.

[161] Cox AD, Der CJ. *Ras history: The saga continues.* Small GTPases 2010; 1: 2-27.

[162] Ismail SA, Chen Y-X, Rusinova A et al. *Arl2-GTP and Arl3-GTP regulate a GDI-like transport system for farnesylated cargo.* Nat. Chem. Biol. 2011; 7: 942-949.

[163] Buss JE, Solski PA, Schaeffer JP, MacDonald MJ, Der CJ. *Activation of the cellular proto-oncogene product p21Ras by addition of a myristylation signal.* Science 1989; 243: 1600-1603.

[164] Brandt-Rauf PW, Carty RP, Chen J, Avitable M, Lubowsky J, Pincus MR. *Structure of the carboxyl terminus of the RAS gene-encoded P21 proteins.* Proc Natl Acad Sci U S A 1988; 85: 5869-5873.

[165] Srinivasan K, Subramanian T, Spielmann HP, Janetopoulos C. *Identification of a farnesol analog as a Ras function inhibitor using both an in vivo Ras activation sensor and a phenotypic screening approach.* Mol Cell Biochem 2014; 387: 177-186.

[166] Pawson T, Warner N. *Oncogenic re-wiring of cellular signaling pathways.* Oncogene 2007; 26: 1268-1275.

[167] Recktenwald C V., Mendler S, Lichtenfels R, Kellner R, Seliger B. *Influence of Ki-ras-driven oncogenic transformation on the protein network of murine fibroblasts.* Proteomics 2007; 7: 385-398.

[168] Bauer B, Mirey G, Vetter IR et al. *Effector recognition by the small GTP-binding proteins Ras and Ral.* J Biol Chem 1999; 274: 17763-17770.

[169] Rojas AM, Fuentes G, Rausell A, Valencia A. *The Ras protein superfamily: Evolutionary tree and role of conserved amino acids.* J. Cell Biol. 2012; 196: 189-201.

[170] Milroy L-G, Ottmann C. *The renaissance of Ras.* ACS Chem Biol 2014; 9: 2447-58.

[171] Cromm PM, Spiegel J, Grossmann TN, Waldmann H. *Direct Modulation of Small GTPase Activity and Function.* Angew Chem Int Ed Engl 2015; 54: 13516-37.

[172] Prakash P, Gorfe AA. *Lessons from computer simulations of Ras proteins in solution and in membrane.* Biochim Biophys Acta 2013; 1830: 5211-8.

[173] Dar AC, Das TK, Shokat KM, Cagan RL. *Chemical genetic discovery of targets and anti-targets for cancer polypharmacology.* Nature 2012; 486: 80-4.

[174] Ostrem JM, Peters U, Sos ML, Wells JA, Shokat KM. *K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions.* Nature 2013; 503: 548-51.

[175] Patgiri A, Yadav KK, Arora PS, Bar-Sagi D. *An orthosteric inhibitor of the Ras-Sos interaction.* Nat Chem Biol 2011; 7: 585-7.

[176] Leshchiner ES, Parkhitko A, Bird GH et al. *Direct inhibition of oncogenic KRAS by hydrocarbon-stapled SOS1 helices.* Proc Natl Acad Sci U S A 2015; 112: 1761-6.

[177] Maurer T, Garrenton LS, Oh A et al. *Small-molecule ligands bind to a distinct pocket in Ras and inhibit SOS-mediated nucleotide exchange activity.* Proc Natl Acad Sci U S A 2012; 109: 5299-304.

[178] Winter JJG, Anderson M, Blades K et al. *Small molecule binding sites on the Ras:SOS complex can be exploited for inhibition of Ras activation.* J Med Chem 2015; 58: 2265-74.

[179] Sun Q, Burke JP, Phan J et al. *Discovery of small molecules that bind to K-Ras and inhibit Sos-mediated activation.* Angew Chem Int Ed Engl 2012; 51: 6140-3.

[180] Yan C, Liu D, Li L et al. *Discovery and characterization of small molecules that target the GTPase Ral.* Nature 2014; 515: 443-7.

[181] Forbes SA, Beare D, Gunasekaran P et al. *COSMIC: exploring the world's knowledge of somatic mutations in human cancer.* Nucleic Acids Res 2014; 43: D805-11.

[182] Zipfel PA, Brady DC, Kashatus DF, Ancrile BD, Tyler DS, Counter CM. *Ral activation promotes melanomagenesis.* Oncogene 2010; 29: 4859-64.

[183] Fernández-Medarde A, Santos E. *Ras in cancer and developmental diseases.* Genes Cancer 2011; 2: 344-58.

[184] Poulikakos PI, Solit DB. *Resistance to MEK inhibitors: should we co-target upstream?* Sci Signal 2011; 4: pe16.

[185] Zhang C, Spevak W, Zhang Y et al. *RAF inhibitors that evade paradoxical MAPK pathway activation.* Nature 2015; 526: 583-6.

[186] *World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013*; 310, 2191-4.

[187] Vallespín, E. et al. *Customized high resolution CGH-array for clinical diagnosis reveals additional genomic imbalances in previous well-defined pathological samples. Am. J. Med. Genet. Part A 2013*; 161, 1950-1960.

[188] Curtis RK, Oresic M, Vidal-Puig A. *Pathways to the analysis of microarray data. Trends Biotechnol 2005*; 23: 429-35.

[189] Kaddi CD, Parry RM, Wang MD. *Multivariate Hypergeometric Similarity Measure. 2013*; 10: 1505-1516.

[190] Kaddi C, Parry RM, Wang MD. *Hypergeometric Similarity Measure for Spatial Analysis in Tissue Imaging Mass Spectrometry. 2011 IEEE Int Conf Bioinforma Biomed 2011*; 604-607.

[191] Palomares, M. et al. *Characterization of a 8q21.11 microdeletion syndrome associated with intellectual disability and a recognizable phenotype. Am. J. Hum. Genet. 2011*; 89, 295-301.

[192] Shaffer, L. G. et al. *The discovery of microdeletion syndromes in the post-genomic era: review of the methodology and characterization of a new 1q41q42 microdeletion syndrome. Genet. Med. 2007*; 9, 607-16 .

[193] Mefford, H. C. et al. *Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. N. Engl. J. Med. 2008*; 359, 1685-99.

[194] Molin, A.-M. et al. *A novel microdeletion syndrome at 3q13.31 characterised by developmental delay, postnatal overgrowth, hypoplastic male genitals, and characteristic facial features. J. Med. Genet. 2012*; 49, 104-9.

[195] Köhler, S. et al. *Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am. J. Hum. Genet. 2009*; 85, 457-64.

[196] Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. *Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet. 2013*;14:681-91.

[197] Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S. *Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. Hum Mutat.* 2012;33:803-8.

[198] Lochmüller H. *Rare diseases need global solutions: new international initiatives in rare disease omics research. Newsletter British Soc Gen Med. 2013*; 1:2-3.

Apéndice

Los documentos incluidos en este capítulo corresponden a otros artículos publicados durante la etapa predoctoral no presentados como aval de la Tesis Doctoral, así como a comunicaciones realizadas en forma de póster en congresos científicos como parte de la difusión de los trabajos realizados.

Publicación: ***Revealing the relationship between human genome regions and pathological phenotypes through network analysis. Bioinformatics and Biomedical Engineering. Lecture Notes in Computer Science. IWBBIO 2017.***

Revealing the Relationship Between Human Genome Regions and Pathological Phenotypes Through Network Analysis

Elena Rojano[1(✉)], Pedro Seoane[1], Anibal Bueno-Amoros[1], James Richard Perkins[2], and Juan Antonio Garcia-Ranea[1,3]

[1] Department of Molecular Biology and Biochemistry, University of Malaga (UMA), 29010 Malaga, Spain
{elenarojano,seoanezonjic,ranea}@uma.es
[2] Research Laboratory, Regional University Hospital of Malaga (IBIMA), 29009 Malaga, Spain
[3] CIBER de Enfermedades Raras, 28029 Madrid, Spain

Abstract. Recent advances in sequencing technologies allow researchers to investigate diseases resulting of genomic variation. This allows us to further develop the concept of precision medicine and determine the best treatment for each patient. We have focused on developing tools for studying genomic loci associated to pathological traits from the perspective of network analysis. We have obtained from DECIPHER database patient information which includes their affected genomic regions by Copy Number Variations (CNV) and their pathologies described as Human Phenotype Ontology phenotypes. We have used different metrics for calculating association values between phenotypes and affected genomic regions to determine which method fits better to our data. The results obtained in this work, can be used in prediction systems for determining and ranking which genomic regions are associated to a concrete phenotype, in order to help clinicians with their diagnosis.

Keywords: Network analysis · Pathological phenotypes · Precision medicine · Rare diseases · CNV

1 Introduction

Many human diseases are due to changes in the genome that affect functional elements such as genes or regulatory elements. Copy Number Variations (CNVs) represent an important class of genetic variation that can affect large areas of the genome. They are caused by duplication or deletion of large genomic regions [1]. There are now many research groups investigating these variants and looking at their association with pathological traits and patient phenotypes, with the aim of better understanding disease and developing personalised therapies [2]. Their study can be aided by the use of systems biology, by creating networks and studying the relationships between their elements [3]. This approach has

I. Rojas and F. Ortuño (Eds.): IWBBIO 2017, Part I, LNBI 10208, pp. 197–207, 2017.
DOI: 10.1007/978-3-319-56148-6_17

Publicación: ***Systematic identification of phenotypically enriched loci using a patient network of genomic disorders. BMC Genomics 2016.***

Reyes-Palomares et al. BMC Genomics (2016) 17:232
DOI 10.1186/s12864-016-2569-6

BMC Genomics

RESEARCH ARTICLE Open Access

Systematic identification of phenotypically enriched loci using a patient network of genomic disorders

Armando Reyes-Palomares[1,2,4*], Aníbal Bueno[1], Rocío Rodríguez-López[1,2], Miguel Ángel Medina[1,2], Francisca Sánchez-Jiménez[1,2], Manuel Corpas[3] and Juan A. G. Ranea[1,2*]

Abstract

Background: Network medicine is a promising new discipline that combines systems biology approaches and network science to understand the complexity of pathological phenotypes. Given the growing availability of personalized genomic and phenotypic profiles, network models offer a robust integrative framework for the analysis of "omics" data, allowing the characterization of the molecular aetiology of pathological processes underpinning genetic diseases.

Methods: Here we make use of patient genomic data to exploit different network-based analyses to study genetic and phenotypic relationships between individuals. For this method, we analyzed a dataset of structural variants and phenotypes for 6,564 patients from the DECIPHER database, which encompasses one of the most comprehensive collections of pathogenic Copy Number Variations (CNVs) and their associated ontology-controlled phenotypes. We developed a computational strategy that identifies clusters of patients in a synthetic patient network according to their genetic overlap and phenotype enrichments.

Results: Many of these clusters of patients represent new genotype-phenotype associations, suggesting the identification of newly discovered phenotypically enriched *loci* (indicative of potential novel syndromes) that are currently absent from reference genomic disorder databases such as ClinVar, OMIM or DECIPHER itself.

Conclusions: We provide a high-resolution map of pathogenic phenotypes associated with their respective significant genomic regions and a new powerful tool for diagnosis of currently uncharacterized mutations leading to deleterious phenotypes and syndromes.

Background

Genomic Structural Variations are one of the main sources of human genome variation. Copy Number Variations (CNVs) naturally occur in the genome of healthy individuals [1, 2], some of them leading to disease [3]. CNVs consist of thousands to millions of bp deletions, duplications, insertions or inversions, recurrent in the population either by inheritance or spontaneous occurrence (*de novo*) [4]. Although the discovery of CNVs was relatively recent, a plethora of genetic association studies have been carried out to understand their evolutionary [5], functional [6] and phenotypic effects [4]. It has been estimated that two genomes can differ approximately about 0.4 % due to CNVs [7] and that these variations have a considerable impact on human health. Several known chromosome imbalances causing complex genomic disorders have been characterized by different medical conditions such as developmental [8, 9], neuropsychiatric [10–12], cancer [13], autoimmune diseases [14] and idiopathic learning disability [15]. However, recent genome wide association studies suggest that the lack of data for individual's medical records is an important limitation to fully understand the genetic basis for many genomic disorders [16, 17]. Initiatives such as the Personal Genomes Project (PGP) [18], Genomics England (http://www.genomicsengland.co.uk/) and the Precision Medicine program [19] aim to provide descriptive records

* Correspondence: armando.reyes@embl.de; ranea@uma.es
[1]Universidad de Málaga, Andalucía Tech, Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, and IBIMA (Biomedical Research Institute of Málaga), E-29071 Málaga, Spain
Full list of author information is available at the end of the article

Publicación: ***PhylomeDB: A database for genome-wide collections of gene phylogenies. Nucleic Acids Research 2008*.**

Published online 25 October 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D491–D496***
doi:10.1093/nar/gkm899

PhylomeDB: a database for genome-wide collections of gene phylogenies

Jaime Huerta-Cepas, Anibal Bueno, Joaquín Dopazo and Toni Gabaldón*

Bioinformatics Department, Centro de Investigación Príncipe Felipe, Avda. Autopista del Saler, 13 Valencia 46013, Spain

Received August 14, 2007; Revised October 3, 2007; Accepted October 4, 2007

ABSTRACT

The complete collection of evolutionary histories of all genes in a genome, also known as phylome, constitutes a valuable source of information. The reconstruction of phylomes has been previously prevented by large demands of time and computer power, but is now feasible thanks to recent developments in computers and algorithms. To provide a publicly available repository of complete phylomes that allows researchers to access and store large-scale phylogenomic analyses, we have developed PhylomeDB. PhylomeDB is a database of complete phylomes derived for different genomes within a specific taxonomic range. All phylomes in the database are built using a high-quality phylogenetic pipeline that includes evolutionary model testing and alignment trimming phases. For each genome, PhylomeDB provides the alignments, phylogentic trees and tree-based orthology predictions for every single encoded protein. The current version of PhylomeDB includes the phylomes of Human, the yeast *Saccharomyces cerevisiae* and the bacterium *Escherichia coli*, comprising a total of 32 289 seed sequences with their corresponding alignments and 172 324 phylogenetic trees. PhylomeDB can be publicly accessed at http://phylomedb.bioinfo.cipf.es

INTRODUCTION

A phylome is defined as the complete collection of phylogenies reconstructed for every single gene encoded in a genome (1). Although the term was coined several years ago, the development of high-quality, genome-wide collections of phylogenetic trees has been previously prevented due to large demands of time and computer power. Only recently, and thanks to new and faster algorithms and computers, the application of phylogenetics to whole genomes has become feasible. Large-scale phylogenetic studies provide very valuable information on the evolutionary relationships between genes of different species (2). Among other applications, the availability of complete phylomes can be exploited to map duplication and speciation events and thus infer orthology relationships (3), to determine the evolutionary relationships among taxa (4) and even to reconstruct ancestral metabolisms (5). Although some databases provide automatically derived and curated phylogenies (6–9), these follow a family-based approach, since they first group the genes into families and subsequently build a single phylogeny for each family. Moreover, the selection of species included is determined by the specific scopes of these databases. PhylomeDB provides phylomes reconstructed following a gene-based approach (3), in which the same high-quality phylogenetic pipeline is applied to each single gene encoded in a given genome. The resulting trees, alignments and tree-based orthology predictions can be easily accessed, queried and downloaded through a user-friendly web interface. In this article, the data content and web features of the first release of PhylomeDB are described.

DATABASE STRUCTURE AND CONTENT

General features

The current version of PhylomeDB contains the phylomes of three relevant organisms, including human and the two model species *Saccharomyces cerevisiae* and *Escherichia coli*. Future releases of PhylomeDB will incorporate phylomes for new species as well as novel versions of existing phylomes that may include different phylogenetic ranges or updated releases of their respective proteomes. To store all data associated with the phylomes, PhylomeDB uses a relational database. For each phylome, PhylomeDB provides: (i) a *feature page*, which contains general information on the proteomes included in the specific phylome as well as all the details of the phylogenetic pipeline used (e.g. http://phylomedb.bioinfo.cipf.es/index.html?Hsapiens001 for Hsapiens001 phylome); and (ii) an individual entry (Figure 1) for each protein encoded in the seed genome that provides the sequences,

*To whom correspondence should be addressed. Tel: +34 963289680; Fax: +34 963289701; Email: tgabaldon@cipf.es

Pósteres

A continuación se adjuntan comunicaciones en forma de póster presentadas en congresos científicos como parte de la difusión de los trabajos realizados.

- 'Discovering the genetic signal underlying cancer cellular heterogeneity in drug repositioning strategy', en *2nd Annual Meeting on Systems Microscopy*, Lovaina. 2013.
- 'RAS superfamily evolutionary expansion in the human protein network interactome', en *Advanced Lecture Course on Systems Biology (SysBio)*, Innsbruck. 2014.
- 'Interaction network distances and connectivity measurements applied to gene priorization in ECM stiffness regulation of breast cancer', en *4th Annual Meeting on Systems Microscopy*, Viena. 2015.
- 'Using phenotype-*loci* network analysis in undiagnosed clinical cases of patients with rare genomic disorders', en X Reunión Anual 2017 CIBERER (Centro de Investigación Biomédica en Red de Enfermedades Raras), Madrid. 2017.

Discovering the genetic signal underlying cancer cellular heterogeneity in drug repositioning strategy

Ian Morilla (1, *), Robert Steininger III (2), Aníbal Bueno (1), Aurelio A. Moya-García (1), Lani Wu (2), Steven Altschuler (2), Juan A.G. Ranea (1)

1. Department of Molecular Biology and Biochemistry, University of Málaga, Málaga (Spain)
2. Green Center for Computational and Systems Biology, Department of Pharmacology, University of Texas Southwestern Medical Center

*Corresponding author: ian.morilla@uma.es - Current Address: Institute of Molecular Life Sciences (MLS), Zurich

Introduction

Resistance to drugs remains an important problem in cancer treatment as, for example, paclitaxel, a prototypic antimitotic drug [1]. Cellular subpopulation composition has been observed related to drug resistance response in some lung tumour colonies [2] pointing that part of the tumour's response to drug could depends on its cellular genetic variability. Gene expression analyses are useful to elucidate some genetic mechanisms used by tumour cells to resist chemotherapeutic drugs. However, current techniques yield averaged data of cell populations missing the effect of cellular heterogeneity in gene expression [2]. In this work we provide an alternative solution for this problem based on a mathematical method of linear multiple response regression, called deconvolution [3] (Fig. I). Our results indicate that a significant genetic expression signal in response to paclitaxel exposure is related to the cellular variability in tumour populations and missed in averaged arrays analysis. We propose that the identification of this genetic signal associated to tumours' cellular composition in the drug response could be key for defining drug repositioning strategies towards personalized medical treatment.

Materials & Methods

Gene Expression Data: A dataset of 15,661 gene expression profiles for NCI-16 tumor cell lines from CellMiner (http://discover.nci.nih.gov/cellminer/home.do). ***Refining the Genes:*** Genes with low expression variation introduce noise into the analysis. We applied three thresholds to remove these genes, obtaining a *refined gene expression matrix* with 567 genes.
Microscopy image-based cellular subpopulations matrix calculation: The matrix with the cellular subpopulation frequency composition of the NCI-16 cell lines (the subpopulation frequency matrix; Fig. IC) was calculated as described in Singh et al, 2010 [2] using systems microscopy approaches on growth tumor cell lines using a multiplexed immunofluorescent marker set (MS: DNA/pAkt/H3K9-Act).

Tumour cell lines selection: *The NCI-60 dataset includes information about the response to paclitaxel treatment of 60 tumour cell lines. The 8 most sensitive and the 8 most resistant lines were selected for this study (Table I).*

	Cell line	Origin	Notation
Sensitive			
Resistant			

Table I. Description of the resistant and sensitive 16 tumorous cell lines selected for this study. *From left column to right: established cell line id in the NCI-60 database; cancer tissue origin; and abbreviated code used in this work.*

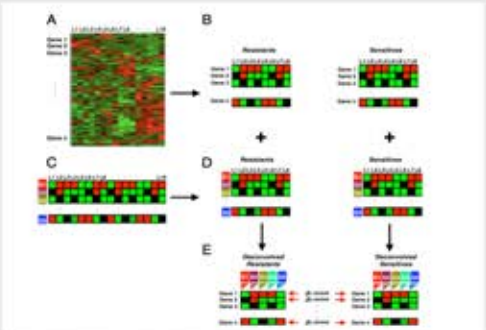


Figure I. Diagram describing the deconvolution process. *Genes expression by cell lines matrix (A); divided up into genes expression by resistant and sensitives cell lines sub-matrices (B); tumours' cellular subpopulation frequency matrix (C); divided up into subpopulations frequency by resistant and sensitives cell lines sub-matrices (D) ; both types of sub-matrices are regressed to calculate the two (resistant and sensitive) intermediate deconvolved matrices (E); finally we calculate the scores for the joint contrast matrix (Δβ-scores).*

Results

Assessing gene refinement and deconvolution processes on the false discovery rate.
Our hypothesis is that genes showing a high variation in their expression levels between resistant and sensitive groups are more likely to be involved in the genetic response to paclitaxel treatment than those other genes that not. In order to compare the performances of the refining and the deconvolution process in the accuracy of detecting gene variation signal by the use of *t*-test and Δβ scores' *p*-value, the analysis of False discovery rate (FDR) was implemented using the method termed SAM (Fig. II). Considering the results of the FDR distribution we can conclude that the gene refining and the deconvolution process applied in tandem significantly increases the performance in detecting genes differentially expressed between paclitaxel resistant and sensitive line.

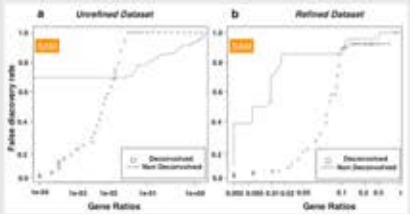


Figure II. *number of genes p-value non deconvolved scores rank (continue lines) and normalised p-value deconvolved scores (circle-lines) for the unrefined gene expression matrices (plot a) and the refined derived matrices (plot b).*

Genetic regulatory signal of tumor cellular heterogeneity on response to paclitaxel.
From the Volcano plot it is possible to see a general correlation between *t*-statistic *p*-values and gene fold change variation in both cases, up and down-regulated genes (Fig. III).

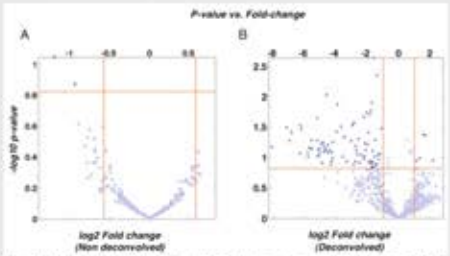


Figure III. *Volcano plots representing gene fold change – x-axis = log2(fold change ratio) and t-test pvalues – y-axis = -log10(p-value) for the non deconvolved (A) and deconvolved (B) datasets. Threshold values used for the selection of genes involved in placlitaxel resistance are indicated with dashed red lines.*

Note that FDRs yield from adjusting these *p*-values in order to control the number of false positives, therefore both approaches (no deconvolved -Fig. IIIA and deconvolved -Fig. IIIB) are comparable.
Clustergrams using different *p*-values and fold change thresholds were built for the deconvolved, the non deconvolved and a random model dataset in order to analyze the effect of these parameters distinguishing between sensitive and resistant lines (data not shown). This supplemental study shows that only by means of the deconvolved approach it is possible to detect a significant subset of genes whose profiles enable the construction of a clustergram that separates well between sensitive and resistant lines (see Fig. IV).

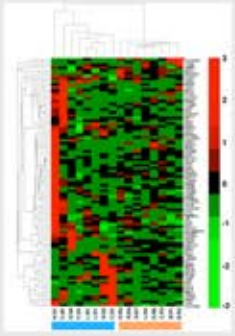


Figure IV. Heatmap of the selected 89 gene profiles within the NCI-16 lines set. *We find out a clear separation between sensitive (blue) and resistant (orange) cell lines.*

Forthcoming

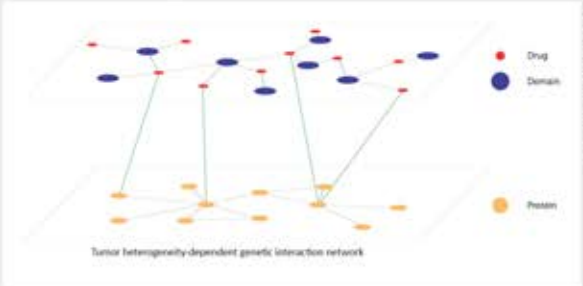


Figure V. We will model different tumor-based gene interaction networks, to consider cellular heterogeneity in cancer. These networks will be the ground of a new structure-based target identification and drug repurposing strategy based in our hypothesis that protein domains can be a major cause of drug polypharmacology (the idea that a single drug can affect multiple targets). Since protein domains are units of structure [4] and there is a limited repertoire of types of domains [5], they are combined to form different proteins with different functions [6]. The reason why a drug bind different protein can be that they share a domain that is the actual target of the drug.

References

[1] Understanding tubulin–Paclitaxel interactions: Mutations that impart Paclitaxel binding to yeast tubulin. Mohan L. Gupta, Jr., Claudia J. Bode, Gunda I. Georg, and Richard H. Himes. 2003. *Proc Natl Acad Sci U S A.* 100(11): 6394–6397.
[2] Patterns of basal signaling heterogeneity can distinguish cellular populations with different drug sensitivities. 2010. Dinesh Kumar Singh, Chin-Jen Ku, Chonlarat Wichaidit, Robert J Steininger III, Lani F Wu and Steven J Altschuler. Molecular Systems Biology 6:369.
[3] Cell type–specific gene expression differences in complex tissues. 2010. Shai S Shen-Orr, Robert Tibshirani, Purvesh Khatri, Dale L Bodian, Frank Staedtler, Nicholas M Perry, Trevor Hastie, Minnie M Sarwal, Mark M Davis & Atul J Butte. Nature Methods. Vol.7 No.4. 287-289.
[4] CATH--a hierarchic classification of protein domain structures. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). Structure, 5(8), 1093–1108
[5] Estimating the number of protein folds and families from complete genome data. Wolf YI, Grishin NV, Koonin EV. J Mol Biol 2000; 299:897–905
[6] Protein domain organisation: adding order. Kummerfeld, S. K., & Teichmann, S. A. (2009). BMC Bioinformatics, 10, 39. doi:10.1186/1471-2105-10-39

RAS Superfamily evolutionary expansion in the human protein network interactome

Aníbal Bueno, Ian Morilla, Aurelio A. Moya-García, Juan A.G. Ranea.*
University of Málaga. Department of Molecular Biology and Biochemistry.
** Corresponding author: anibal@uma.es*

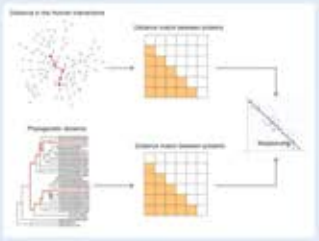
Introduction

Ras proteins are small GTPases involved in cellular signal transduction in eukaryotes, controlling many protein signalling networks related to cellular growth, differentiation and survival. The important functional role of the Ras family in signalling is demonstrated by the fact that 22% of all human tumors contain oncogenic mutations in these proteins.

In this work we examine the relationship between the phylogenetic distance of Ras proteins and their distance in the human protein interactome. In the final part of the work, using the phylogenetic and network distance metrics of Ras pairs, we align and compare Ras sequences in order to search for key positions where residue conservation could lead to the preservation of network localization.

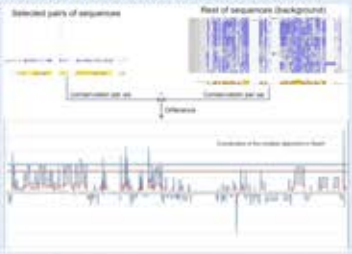
Materials & Methods

General pipeline of the comparison of phylogenetic distance and network interaction distance:



- Phylogenetic trees were obtained from Diez, et al, 2011 [1] and converted into distance matrices.
- Protein interaction networks analyzed:
  - PINA v.20121210 [2].
  - STRING Experimental v9.05 [3].
- Mathematical methods implemented for the distance measurement among protein-protein interaction networks:
  - Laplacian exponential Diffusion Kernel (**DK**): distance = exp (-βL).
  - Commute Time Disffusion Kernel (**CT**): distance = L^+.

Procedure for measuring significant differences in aminoacid conservation:



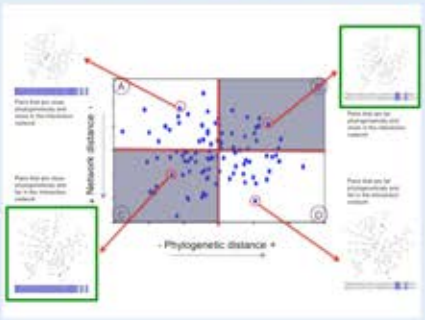
Possible combinations

Different possibilities when comparing phylogenetic distances and protein-protein network distances:

Two different combinations of pairs -out of the initial four- will be interesting to analyze at a sequence level:

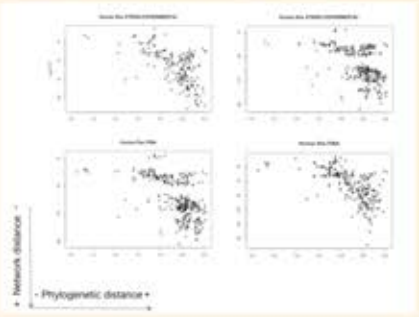
- a) those with a distant phylogenetic connection and a close interaction context.
- b) those with a close phylogenetic relation and a far distance in the network.

In both cases a sequence analysis could lead us to some key residues that makes this happen.



Results & Discussion

As can be seen, there is a tendency between both measures, and there is an interesting group of pairs of proteins in region "B". An analysis of aminoacid conservation was performed between those pairs and the rest of them, so we can establish a relation between that conserved positions and the fact of being in a close interaction context, even with a big phylogenetic divergence.



The sequence comparative analysis shows that, some conserved positions amongst distantly related protein pairs in the phylogenetic tree are mainly involved I Ras protein interaction recognition surfaces , suggesting a potential role of these positions in network localization.

References

1. Diego Diez, Francisca Sánchez-Jiménez, Juan A.G. Ranea (2011) Evolutionary expansion of the Ras switch regulatory module in eukaryotes. Nucleic Acids Res. 2011:gkr154v1-gkr154.
2. Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Makela, T.P. and Hautaniemi, S. (2009) Integrated network analysis platform for protein-protein interactions, Nature methods, 6, 75-77.
3. Jensen et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms Nucleic Acids Res. 2009, 37(Database issue):D412-6.

Interaction network distances and connectivity measurements applied to gene priorization in ECM stiffness regulation of breast cancer

Anibal Bueno (1, *), James R. Perkins (2), Sara Göransson (3), Staffan Strömblad (3), Juan A.G. Ranea (1)

1. Department of Molecular Biology and Biochemistry, University of Málaga, Málaga (Spain)
2. Research Laboratory, IBIMA, Regional University Hospital of Malaga, UMA, Malaga (Spain)
3. Karolinska Institutet. Stockholm. (Sweden)

* Corresponding author: anibal.bueno@uma.es

Introduction

It has been experimentally observed that breast cancer tumour cells change to a malignant phenotype when cultured on a stiffer substrate. Proteomic and transcriptomic experiments were carried out to determine which genes significantly change their expression at the mRNA and protein level as between a high (5000 Pa) and low (400 Pa) stiffness. These experimental approaches returned a high number of genes with significative differential expression but with very divergent functions, or even unknown function. In order to further elucidate the changes we built a systemic approach with the aim of (A) predicting the function/s of those genes and (B) to deduce their implication in the cellular systems that are known to be related with the transduction of mechanosensor signals and the oncogenic signaling pathways.

Materials & Methods

Based on literature [1], public databases [2][3], and expert curation, a set of proteins with a functional role in each of the subsystems within the mechanosensor processes were identified. Those cellular subsystems are: cell-cell adhesion, cytoskeletal regulation, focal adhesion, Hippo signaling pathway, mechanical regulation of nucleus and oncogenic signaling mechanotransduction. The included proteins -in each system- that showed a significative level of differential expression (in proteomic and/or transcriptomic data) between different stiffnesses, were pre-selected.

Figure I. Different subsystems initially stablished for the analysis. *6 gropus of proteins, one for each of the subsystems used for the enrichment analysis.*

Through different models of protein-protein interaction networks and various association metrics, that consider interconnection and network topology, a set of predictors was bulit and validated specifically for each of the aforementioned systems (Fig. II and Fig. III).

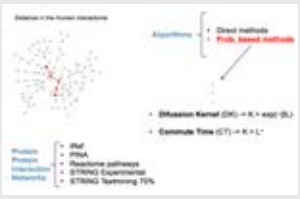


Figure II. Databases and algorithms used. *All the databases used as protein interaction models and the different kinds of metrics proposed.*

Five different models of the interactome, in the form of protein interaction networks, were analyzed: iRef [4], PINA [5], STRING Experimental [6], STRING Textmining [6] and Reactome Pathways [7] (Fig. II).

In relation to the association analysis, obviating the direct methods, two different probabilistic algorithms were used: **Commute Time Kernel** [$K = L+$] and **Exponential Laplacian Diffusion Kernel** [$K = exp(-\beta L)$] due to them being the algorithms that have shown the greatest intrinsic capacity for analysis in the sense of contextual network relationships [8].

For each of the 6 previously described biological systems, associated with the stiffness process, the metric and the network model that showed the best performance as a predictor was selected (using the ROC [Receiver Operating Charasteristic] test to measure that performance using the cross validation method Leave One Out).

Results

Using the predictors that showed a better performance in each subsystem (see Fig. III), we determined which genes with significative differential expression are highly related in the interactome with the ones that we consider as a reference in each set. Those genes selected by our algorithm present a high probability of being directly involved in the molecular mechanosensor mechanism of the particular process that leads to the transformation to a malignant tumoural cell.

Redefining the initial subsystems.

As a result of the ontology database search[2][3] and the data clusterization obtained from the interaction network analysis (Fig. II and Fig. III) we decided to redefine the initial susbsystems in order to be more accurate and to have a better predictive power. Finally our groups were: 1) cell adhesion and 2) cell adherence (both obtained as a subdivision of "cell-cell adherence" after the analysis), 3) cytoskeletal regulation, 4) focal adhesion, 5) hippo pathway, 6) mechanical regulation of nucleus, 7) oncogenic signaling mechanotransduction, 8) regulation of cell proliferation, 9) Wnt receptor signaling, 10) smoothened signaling pathway (8, 9 and 10 are added as a result of a review of the literature and the available public databases).

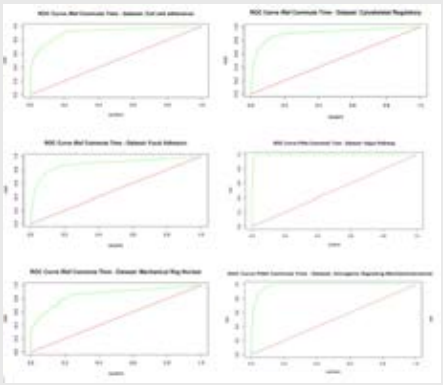


Figure III. Performance of the predictor associated to each subsystem (as a ROC Curve). *After analyzing each system using every network and every algorithm, we choose the best combination of them to be used as a predictor. A Reciever Operating System based on a Leave One Out cross-validation method is used to determine the performance.*

Once the predictor is set up we are able to obtain a prioritized list of proteins for each system with an associated probability of being involved in the molecular process. Those proteins fit the following features (Fig. IV):

1) They are not included in the initial dataset.
2) They show a significant differential expression between the different stiffness media.
3) They are classified as potential new proteins involved in the system by our predictors based on the network analysis (p-value <= 0,05).

Figure IV. Example of protein list *including data about: proteomic and transcriptomic expression, p-value of association to each system according to the predictor (if it is below a p-value of 0,05).*

Forthcoming

We are able, finally, to build networks in order to visualize the relationship between the proteins that originally were included into the systems and the new ones that our predictors suggest that may be implied also in the different functions defined. That way it is possible to see the interactions in context and to decide about further experiments to be carried out with the aim of checking the validity of these gene priorization predictions (Fig. V).

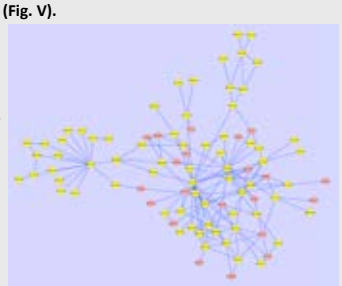


Figure V. Visualization of the network built for the system "cell adhesion" based on the results of the prediction. To obtain the network we take into account the direct relationships in the interactome model used, considering the initial reference set (yellow) and the predicted new proteins (pink). The visualization is carried out using Cytoscape [9].

References

[1] The integrin adhesome: from genes and proteins to human disease. Sabina et. al. Nature Reviews Molecular Cell Biology 15, 273–288 (2014).
[2] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Ashburner et. al. Nat Genet. May 2000;25(1):25-9 (2000).
[3] Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Huang et. al. Nucleic Acids Res. 2009;37(1):1-13. (2009).
[4] iRefIndex: a consolidated protein interaction database with provenance. Razick et. al. BMC Bioinformatics 2008, 9:405. (2008).
[5] Integrated network analysis platform for protein-protein interactions. Wu et. al. Nature methods, 6, 75-77. (2009).
[6] STRING 8--a global view on proteins and their functional interactions in 630 organisms. Jensen et. al. Nucleic Acids Res. 2009, 37(Database issue):D412-6. (2009).
[7] The Reactome pathway knowledgebase. Croft et. al. Nucleic Acids Res. 2014 Jan; 42(Database issue):D472-7. (2014).
[8] Integration of biological data by kernels on graph nodes allows prediction of new genes involved in mitotic chromosome condensation. Heriche et. al. Mol Biol Cell. 2014 Jun 18. pii: mbc.E13-04-0221. (2014).
[9] Cytoscape 2.8: new features for data integration and network visualization. Smoot et. al. Bioinformatics. 2011 Feb 1;27(3):431-2. (2011).

Using phenotype-loci network analysis in undiagnosed clinical cases of patients with rare genomic disorders

Anibal Bueno(1), Rocio Rodriguez-Lopez(1, 2), Armando Reyes-Palomares(3), Manuel Corpas(4), Julian Nevado(2, 5), Pablo Lapunzina(2, 5), Francisca Sanchez-Jimenez(1, 2) & Juan A.G. Ranea(1*, 2).

1 Department of Molecular Biology and Biochemistry, University of Malaga, Malaga, 29071, Spain.
2 CIBER de Enfermedades Raras, Spain.
3 European Molecular Biology Laboratory (EMBL) Heidelberg, Meyerhorfstrasse, 1, 69117, Heidelberg, Germany.
4 The Genome Analysis Centre. Norwich Research Park. Norwich. NR4 7UH UK.
5 Instituto de Genética Médica y Molecular (INGEMM), IdiPAZ; Hospital Universitario La Paz. Madrid, Spain.

*Corresponding author: ranea@uma.es

INTRODUCTION

Copy Number Variations (CNVs) are genomic structural variations (deletions, duplications or translocations) frequently observed in healthy individuals, but they can also lead to disease, being the etiology of many known genetic/genomic disorders. Array-comparative genomic hybridization (aCGH) and single nucleotide polymorphisms arrays (SNParrays) are the main technologies used to interrogate CNVs. These technologies have enabled the identification of novel imbalances in individuals with intellectual disability, autistic disorders and congenital malformations in the past recent years. The continuous use of microarrays has enhanced the development of different freely available databases of CNVs. We used these databases -including the phenotypic information- to create a computational network-based methodology for helping in the diagnosis of new clinical cases [1].

MATERIALS & METHODS

We built tripartite networks with thousands of phenotype-patient-genotype associations using a dataset of 10,324 patients with low prevalent genomic disorders associated to *de novo* CNVs from DECIPHER database [2]. These networks included 14,227 CNVs and 2,583 HPO different phenotypes.

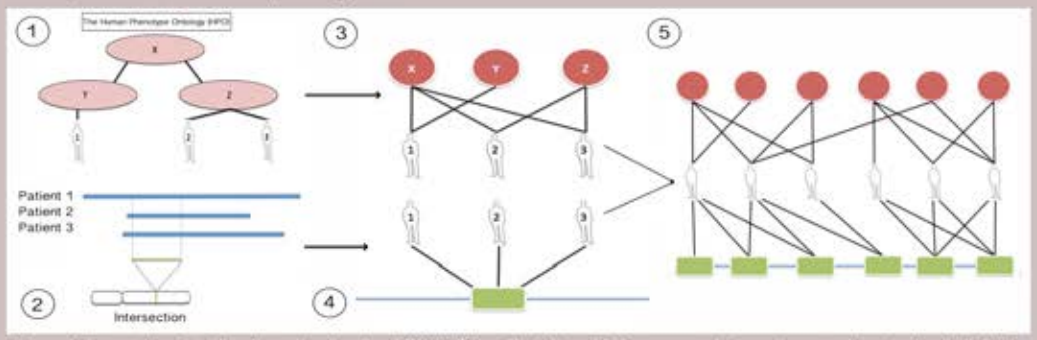


Figure I. *Generation of a tripartite network using DECIPHER patient data. Circles represent phenotypes and rectangles loci. 1) Patients are phenotypically annotated using HPO terms; 2) A locus is generated as the chromosomal region where a set of patient CNVs overlap; 3) The HPOs-patients subnetwork; 4) The patients-loci subnetwork; 5) The final tripartite network.*

We applied the Hypergeometric measure to analyze the relationships between phenotypes and genotypes through patients within the whole network [3].

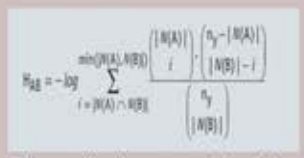


Figure II. *Hypergeometric Index equation.'A' represents a phenotype node and 'B' a locus node within the tripartite network.*

We then aimed to identify new significant pathological associations in mutated regions to assess the potential of this network approach to assist in the diagnosis of novel cases in the genetic clinical routine.

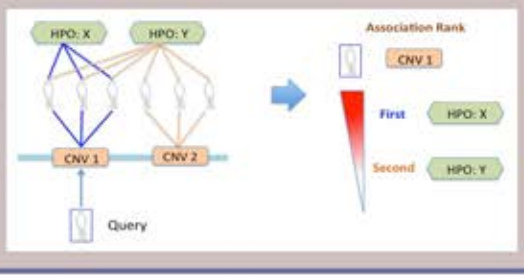


Figure III. *Identification of phenotype-locus associations for new clinical cases. A CNV from a new patient (Query) is assigned to a locus (CNV 1) in the tripartite network by genomic overlap comparison (left side of the figure). All the phenotypes associated to patients are ranked based on their Hyl association value to the query locus (right side).*

RESULTS

The systemic approach implemented in this work is able to better define the relationships between phenotypes and specific *loci*, by exploiting large-scale association networks of phenotypes and genotypes in thousands of patients with rare disorders and complex pathologies. The application of the described methodology has shown a high potential helping in the diagnosis of novel clinical cases, ranking phenotypes by locus specificity and reporting putative new clinical features that may suggest additional clinical follow-ups. The proof of concept developed over a set of novel clinical cases demonstrates that this network based methodology could help in improving the precision of the patients' clinical record and the characterization of rare syndromes.

Patient	Deletion	chr 17	mutation start hg19: 34,911,952	mutation end hg19: 36,510,799	
Observed phenotypes	Associated phenotypes	Hyl rank	Penetrance	% max	Node overlap
Autistic behav. dysplasia (HP:0000717)	+ Abnormality of the kidney (HP:0000077) + Abnormality of the genitourinary system (HP:0000119) + Abnormality of the upper urinary tract (HP:0010935) + Abnormality of the urinary system (HP:0000079)	2.18	50%	71%	100%
		2.87	67%	71%	100%
		3.74	50%	70%	100%
		3.43	56%	66%	100%

Figure IV. Example of the analysis for a novel clinical case. Header: Chromosome coordinates of the CNV. Columns: (1) Observed phenotypes reported by the physician; (2) The list of phenotypes identified by our method. '+' indicates that the HPO term is the most specific one. '->' is used for parental terms; (3) Hyl (Hypergeometric Index) Rank; (4) Penetrance; (5) % max; (6) Node overlap.

REFERENCES

[1] Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth H V, Bailleul-Forestier I et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res 2014; 42: D966–74.

[2] Firth H V., Richards SM, Bevan a. P, Clayton S, Corpas M, Rajan D et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet 2009; 84: 524–533.

[3] Fuxman Bass JI, Diallo A, Nelson J, Soto JM, Myers CL, Walhout AJM. Using networks to measure similarity between genes: association index selection. Nat Methods 2013; 10: 1169–76.

UNIVERSIDAD DE MÁLAGA

ciberer
CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED DE ENFERMEDADES RARAS

Este trabajo se enmarca dentro del área de la Biología de Sistemas, un campo de la Ciencia relativamente nuevo. Más concretamente, los estudios aquí presentados hacen uso de metodologías de modelado y análisis de sistemas biológicos y biomédicos, a nivel molecular, con la finalidad de orientar posibles diseños de fármacos o diagnósticos clínicos en base a los resultados obtenidos mediante técnicas computacionales y análisis estadísticos.

Tres bloques diferenciados sustentan esta Tesis Doctoral:

1) La búsqueda de nuevas proteínas implicadas en procesos patológicos, a través de cálculos estadísticos de distancias en redes funcionales de proteínas y técnicas de validación de las predicciones realizadas.

2) El análisis de la relación entre la evolución molecular y funcional de la familia de proteínas RAS con la finalidad de detectar aquellos componentes estructurales clave para la conservación del contexto de interacciones y así poder orientar el diseño de nuevos fármacos que permitan bloquear las rutas de señalización oncogénicas en esta familia de proteínas.

3) La interpretación conjunta de datos de pacientes con enfermedades raras a nivel fenotípico y genotípico, mediante técnicas de Medicina de Sistemas y análisis de redes a varios niveles, con la intención de obtener relaciones estadísticamente significativas entre las mutaciones genómicas y las características clínicas, de modo que los diagnósticos puedan ser más efectivos.

Los resultados aquí mostrados avalan el uso de estas técnicas bioinformáticas en la investigación biomédica.

UNIVERSIDAD DE MÁLAGA