VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

TESIS DOCTORAL

(RESUMEN EN CASTELLANO)

BIOINFORMÁTICA PARA INTEGRAR INFORMACIÓN DE LA PROTEÍNA Y DEL GEN EN UN CONTEXTO RELACIONAL, APLICACIÓN A LOS DATOS PROTEÓMICOS Y TRANSCRIPTÓMICOS HUMANOS

CONRAD FRIEDRICH DROSTE

DIRECTOR

DR. JAVIER DE LAS RIVAS SANZ

SALAMANCA, 7 DE JULIO DE 2017

Índice

1. **INTRODUCCIÓN**

La célula es un sistema biológico complejo que constituye la unidad básica de todos los seres vivos complejos, siendo el ser humano uno de ellos. La proliferación celular, la diferenciación y las interacciones con el ambiente requieren de la producción, el ensamblaje, la operación y la regulación de miles de componentes. La investigación acerca de cómo las células funcionan a nivel molecular con tal alta precisión es uno de los retos más importantes en la biología molecular moderna (Zhu, Gerstein et al. 2007).

Gracias a los avances técnicos, los investigadores tienen la oportunidad de profundizar en estos sistemas complejos a diferentes niveles celulares: (1) los costes más bajos y la reducción de tiempo en la secuenciación de genomas ha permitido descubrir más información acerca del ADN y sus elementos, como regiones codificantes, marcos abiertos de lectura, elementos reguladores, operones, mutaciones, deleciones, duplicaciones, etc.; (2) el desarrollo de nuevas técnicas transcriptómicas como el RNA-Seq y las mejoras técnicas y el bajo costo de técnicas más usadas como el microarray, ha permitido investigar con mayor profundidad la transcripción en múltiples sistemas biológicos; (3) el siguiente nivel corresponde con el proteoma de la célula, cuyo desarrollo está mejorando notablemente en los últimos años, no sólo desde el punto de vista de la calidad de la identificación de proteínas, sino también en la cantidad de elementos identificables.

Dado que el primer paso hacia una mejor compresión de la célula es el acceso público a todos estos datos, el segundo paso natural corresponde con la interpretación de dichos datos de una forma razonada y reproducible. La Bioinformática es la disciplina encargada de abordar este segundo reto, y, en el caso particular de la Biología de sistemas, la utilización de redes neuronales o redes biológicas es la metodología de representación y análisis más usada. En una red biológica, cada elemento participante en un proceso celular se representará con un nodo, mientras que la relación entre los participantes se representará con una arista o conexión. Las redes pueden representar (i) procesos biológicos simples; (ii) la combinación de varios procesos o rutas; o (iii) directamente un interactoma, el cuál incluiría todo elemento y procesos biológicos llevados a cabo por un organismo.

El proyecto de tesis aquí descrito se centra en el procesado, análisis, integración e interpretación de datos de origen transcriptómico y proteómicos, denominados datos proteogenómicos. Como punto de partida, se han usado datos procedentes de linfocitos B y T o de líneas celulares primarias de pacientes con leucemia.

Uno de los objetivos principales en el campo de la Bioinformática es el desarrollo de nuevas herramientas que faciliten el análisis masivo de bases de datos biológicas, para así generar resultados científicos de los datos ya existentes. De entre ellas, el meta-análisis consiste en analizar sets de datos diferentes para conseguir resultados nuevos y robustos. El interactoma humano que conocemos hoy en día está generado siguiendo está filosofía, y comprende alrededor de 25000 proteínas, 1000 metabolitos, multitud de elementos proteicos desconocidos y ARN funcionales (Barabasi, Gulbahce et al. 2011).

La disposición de las entidades biológicas de un organismo en una única red biológica o interactoma permite analizar la misma en búsqueda de aquellos elementos de vital relevancia o

claves. Estos nodos claves son los responsables en muchas ocasiones de enfermedades, como el cáncer, cuando su función se ve alterada. Por tanto, su posición en dicha red es capaz de revelar en qué procesos está implicado, y así valorar posibles tratamientos con ellos como dianas terapéuticas. Sin embargo, analizar un interactoma completo se antoja prácticamente imposible hoy en día. Es necesario reducir su complejidad, analizando partes del mismo relacionadas entre sí, y que sean de interés según qué contexto biológico se quiera estudiar. En este sentido, se antoja necesaria toda información relevante que generen otras disciplinas como la genética, la biología celular o la biofísica.

Gran parte de la comunidad científica dedicada al estudio de la Biología de sistemas presta especial atención a la investigación de las interacciones proteína-proteína, usando modernas técnicas como el sistema doble híbrido (Suter, Kittanakom et al. 2008) o la espectrometría de masas. Con ellas, podemos verificar: (i) publicaciones ya existentes sobre interacción de proteínas; (ii) relaciones entre proteínas predichas a partir de un organismo modelo; y (iii) nuevas interacciones de proteínas (Rual, Venkatesan et al. 2005). Estas redes de interacciones proteína-proteína constituyen la mayor y más diversa fuente de datos disponibles (Zhu, Gerstein et al. 2007) para la generación de redes biológicas. El análisis de dichas redes se considera, en muchos de los casos, una representación fidedigna de la realidad, así como las proteínas clave o centrales.

La integración de datos biológicos de origen proteogenómicos con una red biológica creada a partir de bases de datos ya existentes es uno de los principales objetivos de este proyecto de tesis. Así, aquí se detalla cómo funciona Path2enet, una herramienta bioinformática desarrollada con el objetivo de mapear datos biológicos en redes biológicas para definir qué proteínas son clave en un proceso o tejido biológico concreto.

1.1 Estado del arte de la proteogenómica

Cualquier análisis proteogenómico ha de incluir datos genómicos (secuenciación del ADN y ESTs) y transcriptómicos (secuenciación del ARNm, ARNmi y del ARNr), usados para generar bases de datos específicas que ayudan a la interpretación de los datos proteómicos (LC-MS o MS). De hecho, los datos proteómicos validan y refinan los niveles de expresión de genes obtenidos de los datos transcriptómicos de cualquier modelo génico.

1.1.1 Datos de expresión génica

1.1.1.1 ESTs

Los marcadores de secuencia expresada o ESTs representan la fuente de datos más antigua utilizada para la realización de esta Tesis Doctoral, y son cortas lecturas de ADN, que representan los genes expresados en determinado tejido o tipo celular. Ya fueron usados en uno de los primeros análisis de expresión génica en 1994 (Fields C., 1994), y en este proyecto se han incorporado información de la base de datos Unigene.

1.1.1.2 Microarray o chip de ADN

El microarray o chip de ADN constituye la técnica de medida de expresión génica más comúnmente utilizada en las últimas dos décadas, tanto en investigación como en Biomedicina. De entre las plataformas comerciales existentes, Affymetrix es la más extendida (https://www.affymetrix.com/site/mainPage.affx). Esta tecnología ha contribuido enormemente al desarrollo de la investigación biomédica, proporcionando lecturas fiables y reproducibles de expresión génica a bajo costo.

1.1.1.3 Secuenciación de ARN

La secuenciación de ARN (con siglas en inglés RNA-Seq) es la técnica actual de secuenciación más extendida, y cuyo uso está reemplazando, en muchas ocasiones, a los chips de ADN. Esta tecnología usa técnicas de secuenciación masivas para determinar la abundancia relativa de ADNc en un tejido o tipo celular concreto. Varias son las ventajas que proporciona esta tecnología de secuenciación: (1) los investigadores pueden usar todas las tecnologías de secuenciación de ADN; (2) la amplificación del fragmento de ADNc no es necesaria; (3) las lecturas están mapeadas directamente a la base de datos de referencia; (4) se puede crear un genoma si no existe ninguno de referencia; (5) se obtiene el nivel de expresión y la estructura/secuencia en un único experimento; (6) hay baja señal de fondo; y (7) requiere poca cantidad de muestra de ADNc.

Sin embargo, tiene algunas limitaciones: (1) las librerías de ADNc no permiten analizar todos los tipos de transcritos; (2) la fragmentación de largas moléculas de ARN es necesaria; (3) los fragmentos pequeños pueden ser idénticos; (4) limitaciones a la hora de analizar los datos con herramientas Bioinformáticas; y (5) gran cantidad de los transcritos son costosos pero interesantes de analizar desde el punto de vista biológico (Wang, Gerstein, & Snyder, 2009).

Path2enet incluye dos sets de datos de RNA-Seq a priori procesados como referencia, ambos con la señal de expresión normalizada a FPKMs (Fragment Per Kilobase of Exon per Million): (i) set de datos de Human Body Map 2.0, incluyendo muestras de 16 tejidos con 3 réplicas; y (ii) el set de datos de The Human Protein Atlas. Estos dos sets de datos constituirán la referencia para el análisis de los datos transcriptómicos obtenidos mediante RNA-Seq para la línea celular de células B de Ramos (experimento SRX105534: http://www.ncbi.nlm.nih.gov/sra/SRX105534).

1.1.2 Datos proteómicos

1.1.2.1 Espectrometría de Masas

La espectrometría de masas es una técnica proteómica indispensable en la investigación biomolecular y bioquímica. Esta técnica permite determinar interacciones proteína-proteína así como concentración proteica en muestras biológicas, y origen subcelular de las mismas. Un experimento típico consta de 5 pasos: (i) aislamiento proteico mediante fraccionamiento o selección por afinidad con gel dimensional electroforético; (ii) degradación de las proteínas a péptidos; (iii) separación de los péptidos por cromatografía líquida de alta presión o LC; (iv) elución para crear un espectro de masas; y (v) espectrometría de masas en tándem para obtener una lista de péptidos candidatos fiable.

1.1.2.2 Anticuerpos

SEC-MAP es una técnica proteómica basada en la afinidad de anticuerpos hacia ciertas proteínas preseleccionadas con anterioridad. Esta técnica ha sido empleada con éxito para detectar las proteínas existentes en una única muestra, así como información acerca del tamaño y la localización subcelular. La figura 5 muestra un ejemplo de experimento con SEC-MAP.

1.2 Estado del arte del análisis de redes biológicas

1.2.1 Teoría de grafos

Las redes biológicas son muy grandes y han de ser interpretadas estadística y visualmente de forma que podamos adentrarnos en su significancia biológica sin perder perspectiva (Barabasi, Gulbahce et al. 2011). Por ello, estas redes son normalmente interpretadas como grafos, y como tales, toda la teoría de grafos existente puede ser aplicada a ella con ciertas consideraciones. Consideramos las proteínas como nodos y las interacciones como aristas. De esta forma, las interacciones físicas entre proteínas (proteína-proteína) que deseamos usar y estudiar en esta Tesis Doctoral, pueden ser interpretadas como un grafo indirecto (sin dirección). La información biológica relevante podrá deducirse mediante diferentes análisis sobre las redes, teniendo en cuenta conceptos sencillos como el camino más corto entre dos nodos/proteínas, la distancia de dicho camino, o el diámetro del mismo (detallado en la Tesis).

1.2.2 Set de datos de redes biológicas

1.2.2.1 Redes de interacción proteína-proteína

A la hora de general redes de interacción de proteínas, uno de los principales aspectos a tener en cuenta es la procedencia de la información. La base de datos de referencia es UniProtKB (Universal Protein KnowledgeBase) y proporciona una amplia información, de libre acceso y de calidad, acerca de las proteínas conocidas de cualquier organismo vivo. Dentro de la misma existen dos sets de datos según su nivel de revisión/curado: SwissProt (altamente curado) y TrEMBL (gran cantidad de información, pero de peor calidad).

En lo que respecta a bases de datos fiables sobre interacción de proteínas, encontramos dos tipos: (A) bases de datos primarias, cuya información está extraída directamente de publicaciones científicas de calidad, y dónde los “curators” o revisores expertos evalúan, interpretan y compilan toda la información; y (B) bases de meta-datos, como por ejemplo “Agile Protein Iteraction DataAnalyzer (APID), utilizada en esta Tesis Doctoral. APID unifica bases de datos primarias sobre interacción de proteínas, como hacen otras (iRefWeb, Mentha, GeneMania y STRING).

Por último, otra fuente de datos acerca de interacción proteína-proteína es la basada en información estructural 3D. Está fuente de datos se sirve de experimentos con espectroscopía de resonancia magnética, microscopía electrónica o cristalografía y difracción con rayos X. La información está compilada en la base de datos Protein Data Bank (PDB).

1.2.2.2 Rutas biológicas

Otra gran fuente de datos curados son las bases de datos sobre rutas biológicas como KEGG, ByoCyc, Pathway Commons y Reactome. Todas ellas almacenan datos sobre procesos biológicos

relevantes a nivel génico, proteómico y metabólico. Así, las rutas biológicas pueden clasificarse, a grandes rasgos, en rutas metabólicas o no metabólicas, siendo estas segundas las que podrían considerarse como redes generalizadas de interacciones proteína-proteína. Además, es esencial la comprensión de la regulación que tiene cada ruta biológica dentro de un organismo complejo. Enfermedades como el cáncer resultan de la desregulación de rutas concretas, y, por esta razón, estas bases de datos incluyen rutas biológicas que reflejan dichas alteraciones respecto de la original, denominadas bajo el nombre de la enfermedad subyacente (Melanoma, Leucemia Mieloide Aguda, Cáncer de Próstata, etc.).

1.2.3 Anotación y enriquecimiento funcional

1.2.4 Herramientas bioinformáticas disponibles

Dado que la herramienta Path2enet aquí desarrollada ha sido escrita en el lenguaje de programación R, vamos a mencionar brevemente las herramientas desarrolladas en R que trabajan con bases de datos de interacción de proteínas o rutas biológicas.

En lo referente al análisis e integración de bases de datos sobre rutas biológicas como KEGG, librerías de R como KEGG.db, Gene2Pathway, KEGGgraph, PaxtoolsR, rBiopaxParser, PathView, o MetaboSignal integran diferentes niveles de información y los utilizan para generar rutas biológicas, redes complejas con varias rutas o completar con nueva información de otras bases de datos.

En el caso de análisis de redes biológicas o de la incorporación de datos más complejos como series temporales, BioNet y pwOmics son algunas de las más relevantes, respectivamente. Además, la librería graphite nos permite acceder libremente a gran cantidad de bases de datos sobre redes biológicas desde R, así como también permite asociar fenotipos o características de muestras que deseemos estudiar a determinadas regiones de una red, mediante el análisis topológico de la red y la teoría de descomposición.

Por último, y centrado en el enriquecimiento funcional basado en rutas biológicas, encontramos librerías en R como SPIA, DEGraph, TopologyGSA, TAPPA, PRS y PWEA. A destacar, SubpathwayMiner, una herramienta para identificar rutas relevantes en una lista de genes y generar subredes a partir de ellas mediante una serie de parámetros, y FGNet, una librería desarrollada en R con interfaz gráfica de ventanas, que permite crear redes funcionales a partir de enriquecimientos biológicos basados en herramientas muy extendidas como DAVID, GeneTerm Linker, gage y topGO.

2. OBJETIVOS

2.1 Problema actual e hipótesis

Actualmente, la gran mayoría de estudios biomoleculares, especialmente en muestras humanas, que están publicados en revistas científicas de alta calidad están centrados en el análisis de una o un pequeño grupo de proteínas, obviando aquellas otras biomoléculas con las que interacciona(n). Esta simplificación inductiva es intrínseca al método científico, pero está en contra del, cada vez más instaurado, concepto de producción de datos ómicos globales. Pensamos que esta frecuente omisión del contexto biológico podría ser parcialmente solucionada con la aplicación de herramientas bioinformáticas y la biología computacional a estudios biológicos.

2.2 Objetivos

El objetivo general de esta Tesis Doctoral es el desarrollo y la aplicación de algoritmos bioinformáticos y métodos para la integración, análisis y visualización de varias fuentes de datos genómicos y proteómicos, así como información biomolecular del organismo humano. El manejo y el análisis integrativo de los datos complejos que están actualmente disponibles para proteínas y genes a nivel global, es un reto para el campo de la Biomedicina y un escenario propicio en el cuál aplicar y desarrollar nuevas herramientas y métodos bioinformáticos. Dentro de este marco de trabajo, la presente Tesis Doctoral desarrollo algunos métodos específicos para integrar varias capas de información biológica y experimental, con el fin de generar redes biomoleculares y facilitar el análisis y la visión global de los procesos biomoleculares relevantes en un contexto biológico preciso. Estos métodos son aplicados, como casos específicos de estudio, a varios sets de datos experimentales: (i) linfocitos humanos (células B y T); (ii) datos proteómicos y transcriptómicos obtenidos de varias líneas celulares de linfoma humano; y (iii) datos proteómicos obtenidos de células aisladas de pacientes con linfoma.

Para ser más preciso, se proponen los siguientes cuatro objetivos específicos:

Objetivo 1: El diseño y desarrollo de una base de datos biológicos integrativa que combine información sobre rutas biológicas, proteínas y expresión. Esta base de datos incluirá datos de tres tipos: primero, información sobre rutas biológicas humanas (obtenida de la base de datos KEGG) combinada con la correspondiente información sobre cada proteína (obtenida de la base de datos UniProt); segundo, datos sobre interacción directa proteína-proteína (obtenida de la base de datos APID); tercero, datos de expresión génica en humanos en diferentes tipos celulares, tejidos y órganos (obtenida de varias fuentes de datos transcriptómicos).

Objetivo 2: El desarrollo de una herramienta o aplicación bioinformática que genere la traducción de rutas biológicas a redes de proteínas, integrando varias capas de información biomolecular en una única vista. La herramienta usa la base de datos biológica integrativa producida en el primer objetivo, transformando las rutas biológicas en redes y enriqueciendo las redes con datos experimentales de interacción proteica y de expresión génica. La herramienta también analiza la expresión génica para determinar si cada gen/proteína de la red biológica está activo (ON) o inactivo (OFF) en un contexto biológico o tipo de muestra específico. De esta forma, queremos reducir la complejidad de las redes y determinar las proteínas activas

bajo condiciones específicas. En general, la herramienta bioinformática será diseñada para analizar y visualizar una o varias rutas biológicas en forma de redes de expresión biológica. Como caso de estudio, hemos aplicado esta herramienta a la investigación de varios tipos celulares específicos humanos.

Objetivo 3: El desarrollo y la aplicación de un análisis integrativo de calidad de set de datos proteómicos (obtenidos por espectrometría de masas o identificación específica por anticuerpos) y set de datos transcriptómicos (obtenidos por expresión génica con microarrays de ARNm y secuenciación de ARNm). Este enfoque proteogenómico quiere proveer una visión global sobre qué proteínas o genes están activados en un contexto determinado y qué funciones biológicas desempeñan. Como caso de estudio, hemos aplicado esta aproximación a un set de datos de células del linfoma de Burkitt.

Objetivo 4: El desarrollo y la aplicación de un procedimiento para integrar datos proteómicos y fosfoproteómicos cuantitativos (obtenidos por espectrometría de masas) y calcular los niveles relativos de expresión de proteínas, así como, mapear las proteínas identificadas a procesos biológicos y rutas de señalización. Como caso de estudio, hemos aplicado este método a un set de datos de pacientes con linfoma en diferentes estadios clínicos.

Todos los métodos aquí presentados han sido probados usando muestras humanas aisladas de diferentes tejidos, estadios de desarrollo, líneas celulares o pacientes. Todos los sets de datos están disponibles con libre acceso público, y, en varios casos, provienen de trabajo experimental en colaboración con el laboratorio número 11 del Centro de Investigación del Cáncer (CiC-IBMCC de USAL/CSIC, Salamanca).

3. RESULTADOS

3.1. Procesado de sets de datos biológicos para construir redes biomoleculares

La herramienta Path2enet usa set de datos de redes neuronales y datos transcriptómicos pre-procesados para generar estas redes biomoleculares.

3.1.1. Conversión de identificadores

Con el fin de establecer un único identificador (ID) por cada proteína, en este trabajo se ha elegido el identificador UniProt (UniProt ID) debido a que: (i) UniProt provee tablas de mapeo y herramientas de alta calidad; (ii) también muestra si una la información de una proteína está revisada por expertos; (iii) es el mismo identificador usado por la base de datos APID; y (iv) es un identificador extendido en la investigación proteómica.

3.1.1.1. Generar la tabla de mapeo del identificador necesita KeggXML2SQLDatabase y un set de datos de EST

La librería escrita en R, *Path2enet*, genera una tabla de mapeo (mediante la función *IDmappingFunction*) para traducir de identificadores ENTREZ Gene o KEGG a identificadores Uniprot. Para ello, se sirve del servidor FTP de UniProt y de su última versión. La tabla de referencia creada es muy importante porque relaciona los IDs de KEGG con los correspondientes IDs de UniProt, y sin ella, no sería posible generar las bases de datos integrativas con estos IDs primarios comunes.

Para generar dicha tabla, únicamente tenemos que proporcionarle una carpeta temporal y el organismo de estudio (ARATH, CAEEL, CHICK, DANRE, DICDI, DROME, ECOLI, HUMAN, MOUSE, RAT, SCHIPO o YEAST) a la función *IDmappingFunction* en R. Sin embargo, la librería Path2enet ya incluye de partida una tabla *IDMappingTable* para humanos, por lo que podemos evitar este paso.

3.1.1.2. Generar la tabla IDmappingTable para sets de datos transcriptómicos

Para el mapeo de datos procedentes de secuenciación de ARNm o de microarrays de ARNm, se necesitan IDs de otras plataformas como ENSEMBL o las plataformas de *Affymetrix* HGU133A o HGU133Plus2.0 (ambas plataformas soportadas por el algoritmo *Barcode*). Además, debido a la ambigüedad de los IDs de probesets, Path2enet los eliminará usando CDFs procedentes de la herramienta *BrainArray* (versión 20).

En este siguiente paso, Path2enet traduce los IDs de Ensemble a los IDs de UniProt/SwissProt para incluir, únicamente, aquellas proteínas revisadas. Es necesario que el usuario descargue previamente el archivo uniprot_sprot.fasta.gz del servidor FTP de UniProt. Para aquellos casos dónde se quiera mapear directamente desde las plataformas de *Affymetrix* HGU133A y HGU133Plus2, la herramienta ya incluye ambas tablas de referencia.

3.1.2. Preparando las bases de datos de redes para Path2enet

Path2enet usa las bases de datos KEGG y Agile Protein Data AnalyzerData para generar sus redes biológicas. Este apartado explica cómo deben prepararse dicha integración.

3.1.2.1. Preparando la base de datos de MySQL

Sirviéndose de MySQL, Path2enet construye una base de datos relacional para almacenar la información de KEGG y APID y acceder a ella fácilmente después. Para ello, es necesario tener instalado SQL en nuestro ordenador y configurar correctamente MySQL (http://www.mysql.com/). Además, necesitamos tener privilegios de usuario para crear y acceder al servidor de MySQL.

3.1.2.2. Preparando APID

Path2enet utiliza la base de datos APID para acceder a información sobre interacción de proteínas experimentalmente probadas. La función *Apid2Sql* en la librería en R permite transferir dicha base de datos a nuestra base de datos relacional en MySQL. Para ello, es necesario que el usuario descargue el interactoma humano previamente del servidor web de APID (http://apid.dep.usal.es).

3.1.2.3. Preparando KEGG

Uno de los objetivos de la herramienta *Path2enet* es almacenar los datos procedentes de archivos KGML de KEGG en una base relacional del tipo MySQL. Para facilitar el uso de esta base de datos, *Path2enet* pone a disposición del usuario una sencilla función en R para seleccionar rutas biológicas concretas de KEGG a partir de los nombres de genes o de las rutas.

Una de las complicaciones a salvar fue la dependencia de *Path2enet* de otras librerías ya implementadas en R. En ese sentido, se creó una función específica para leer e implementar archivos tipo KGML. Esta función lee los archivos, mapea los genes al ID de UniProt correspondiente y genera tablas con la información, base necesaria para introducir en MySQL.

3.1.2.3.1. Creación de la base de datos KEGG

La función *KeggXML2SqlDatabase* es la encargada de llevar a cabo este proceso. Es necesario que el usuario descargue el archivo tipo Brite (br8901.keg de la web http://www.kegg.jp/kegg-bin/get_htext?br08901.keg website), el cual contiene la jerarquía de la base de datos KEGG (categorías A, B o C, de menor a mayor generalidad en los términos).

- Lector de archivos KGML KeggXML2SqlDatabase

La primera parte de la función *KeggXML2SqlDatabase* trata de convertir los archivos KGML a tablas ordenadas con información acerca de los diferentes identificadores de KEGG (Tabla 3). Después, trabajaremos con tablas intermedias para convertir aquellos IDs que lo necesiten a nuestro ID canónico de UniProt.

- Construcción de la base de datos con KeggXML2SqlDatabase

Una vez hemos generado las tablas de mapeo de identificadores tras leer los archivos KGML, pasamos a generar la base de datos MySQL. Para ello, *KeggXML2SqlDatabase* llama a la librería

de R específica para SQL, llamada RMySQL, así podemos acceder, modificar y transferir datos completos desde MySQL al entorno de R y viceversa.

Todas las tablas aquí utilizadas están indexadas para obtener mayor rendimiento. Hemos elegido el entorno de trabajo de MySQL debido a su capacidad para almacenar, procesar e interrogar grandes volúmenes de información.

- Estructura de la base de datos

Así, tras estos pasos, la base de datos va a estar constituida principalmente por dos tablas referencia, originalmente denominadas KEGGDatabase (información sobre la entrada y archivo KGML correspondiente) y KEGGRelation (con información referente al campo Relation de la ficha KGML correspondiente), que pasan a tener un prefijo "Path2enet_" delante tras la conversión. A la primera tabla le hemos añadido información, no presente en KEGG, acerca del número de parálogos encontrados para un mismo gen/proteína, basándonos en el número de identificadores diferentes encontrados para dicho gen/proteína bajo el término *NrParalogous*. La tabla final generada, llamada *Relations*, es una combinación de estas dos tablas.

Usando la jerarquía impuesta por el archivo Brite (KeggBR) descargado anteriormente, *Path2enet* genera meta-rutas específicas durante el proceso de generación de la base de datos MySQL, y las marca con el prefijo mencionado arriba. Un dato relevante, es que tras comparar las rutas biológicas relacionadas con cáncer con aquellas que podemos considerar normales, tenemos un número de entradas muchísimo menor (en número de nodos o genes/proteínas/metabolitos) de elementos biológicos participantes.

- Acceso a la base de datos vía R

La función *searchFunction* en R está diseñada para aquellas personas no familiarizadas con la sintaxis de SQL y que deseen profundizar e interrogar a la base de datos de MySQL. Los parámetros y su explicación vienen detallados en la Tabla 6, así como varios ejemplos de uso a continuación.

3.1.3. Procesado de set de datos transcriptómicos

Las bases de datos integradas en nuestro MySQL no son específicas de tejido o de condición de estudio, aunque muchos de los estudios realizados en investigación se centran en la comparación de dos condiciones o estadios diferentes. Por ello, consideramos implementar datos de expresión génica de diferentes fuentes en nuestra base de datos, y generar así un mapa transcriptómico de referencia.

3.1.3.1. EST

La base de datos UniGene almacena información sobre Marcadores de Secuencia Expresada o ESTs (por sus siglas en inglés). Estos datos dan una idea general de la expresión génica en diferentes tejidos y tipos celulares humanos. Así, la función *mkESTdb* implementada en nuestra librería genera la base de datos a partir de UniGene y la implementa a nuestro MySQL para Homo sapiens. Posteriormente, la función *createEstByTissue* permite seleccionar qué condiciones o tejidos han de incluirse en la base de datos relacional. De forma adicional, la propia librería *Path2enet* incluye una base de datos de ESTs (descripción en la viñeta de ayuda de la librería) con 18880 genes/proteínas en 51 tejidos humanos diferentes.

3.1.3.2. Chip de ADN – Expresión génica con algoritmo Barcode

Path2enet incluye un amplio repositorio de datos de expresión génica, obtenido mediante la inclusión del algoritmo *Barcode*. Este método fue diseñado para determinar los niveles de señal mínima necesaria para asegurar que un gen se encuentra ON ó OFF en un tejido determinado. Así, sus sets de datos (McCall et al., 2011; McCall et al., 2014) incluyen información sobre 17268 genes/proteínas y 195 tejidos y líneas celulares. Toda esta información está incluida en *Path2enet* y los identificadores han sido traducidos a IDs de UniProt.

3.1.3.3. Secuenciación de ARN

Path2enet también incluye dos sets de datos de secuenciación de ARN: (i) Human Body Map 2.0, con información sobre 18744 genes/proteínas y 16 tejidos humanos, con IDs traducidos a IDs de UniProt; y (ii) Human Protein Atlas, con datos sobre 19078 genes/proteínas en 33 tejidos humanos.

3.2. Path2enet para generar, analizar y visualizar las rutas que dirigen redes biomoleculares

3.2.1. Generando las redes biomoleculares

El meta-análisis de APID, KEGG y los datos transcriptómicos sólo es posible si Path2enet es capaz de acceder a todos estos datos de forma eficiente mediante MySQL, para así combinar todos ellos eficazmente. La conversión de los genes/proteínas incluidos en todos estos datos a IDs de UniProt nos permite acceder fácilmente a ellos inequívocamente. La función path2enet en la librería (parámetros en Tabla 9) permite la integración y el análisis de rutas o networks sub-seleccionados a lo largo de estas tres bases de datos. Esta función crea dos objetos “lista” de R que incluyen las redes de proteínas creadas en formato de la librería *igraph* de R, y en formato *data.frame* para su fácil acceso.

3.2.1.1. Grafos de la función path2enet

La función *path2enet* genera tres diferentes grafos y los compara resaltando las interacciones compartidas por los tres: (i) una red con información de interacción proteína-proteína de las bases de datos seleccionadas (por ejemplo, KEGG); (ii) un grafo local pequeño creado de las interacciones obtenidas de la base de datos de rutas biológicas seleccionada (por ejemplo, APID); y (iii) un grafo global más amplio a partir de rutas biológicas de la base de datos seleccionada.

3.2.1.2. Atributos de los grafos de la función path2enet

La información más relevante acerca de los networks está incluida en los slots *vertex atributtes* y *edge attributes*. Cabe recordar que un *vertex* corresponde a un nodo o gen/proteína y un *edge* corresponde con una interacción entre nodos. Existen varias funciones de la librería *igraph* de R que permiten obtener la información de estos slots, así como analizar sus características globales. Alguna de la información de *edges* incluye datos sobre en qué tejidos están relacionados dichos elementos o el tipo de relación que existe (activación, fosforilación, represión, etc.).

3.2.1.2.1. Visualización de grafos con Path2enet

La creación de grafos con formato *igraph* permite que el resto de la comunidad científica pueda acceder y modificar fácilmente dichas redes, ya que esta librería es la más extendida en el uso de grafos. Sin embargo, debido a ciertas limitaciones en las funciones de representación de la librería *igraph* y de la librería *tkplotter* (alternativa para dibujar grafos), nuestra herramienta *Path2enet* incluye dos funciones en R llamadas *graphTKplotterPATHW* y *graphTKplotterTissue*, descritas brevemente a continuación.

3.2.1.2.2. Representación común de grafos con graphTKplotterPATHW

La forma básica de representar nuestras redes o grafos es mediante el uso de esta función de *Path2enet* (Figura 9). Los atributos relacionados con el tipo de interacción entre dos nodos (activación, inhibición, expresión, fosforilación y otros tipos) están marcados en diferentes colores, mientras que si no existe relación definida son interacciones de color negro. Así, los nombres de los nodos corresponden con los IDs de UniProt, si bien estos no muestran el prefijo de organismo ("RAS_HUMAN" pasa a "RAS" si vienen de SwissProt, o "Q7Z6C_HUMAN" a "Q7ZC6" si vienen de TrEMBL) en los plots. Así, los colores de los nodos reflejarán si la proteína está revisada o no en UniProt o si son miembros de una comunidad dentro de la red, mientras que los colores de las interacciones reflejaran el tipo de interacción y cuáles de ellas constituyen el grafo básico.

3.2.1.2.3. Representación tejido-específica de grafos

Ya que los atributos de los nodos generados por la función R *graphTKplotterTissue* incluyen los niveles de EST, chip de ADN y secuenciación de ARN, para diferentes tejidos humanos procedentes de la base de datos UniGene, Gene Expression Barcode, ProteinAtlas y Human Body Map 2.0, podemos implementar esta capa de información en las redes que generemos en base a APID o KEGG. Así, esta función utiliza dicha información para marcar en amarillo aquellos nodos que tienen al menos un transcrito en el tejido seleccionado, en blanco o eliminar aquellos que no poseen ningún transcrito, y marca en azul aquellos de los que no se posee información alguna.

En la Figura 11, se muestra un claro ejemplo de la funcionalidad de incorporar estos datos transcriptómicos, ESTs en este ejemplo, a las redes. Mientras las bases originales APID o KEGG no permiten diferenciar entre subredes internas, con *Path2enet* conseguimos separar claramente entre los elementos Notch y SUH de la ruta biológica con nombre "Notch Signaling Pathway" de KEGG. Los datos de EST aquí incluidos corresponden con hígado, y podemos decir que este grafo es tejido específico, que proporciona una buena visión sobre qué proteínas están interaccionando en hígado dentro de esta ruta, y que es menos complejo al eliminar aquellos genes/proteínas sin EST en hígado. Cualquier otro tejido incluido en UniGene puede ser representado.

En lo referente a los grafos locales y globales generados por la función *path2enet*, compararlos entre sí puede ser de gran utilidad para el usuario para conocer las relaciones intrínsecas de una ruta en un tejido ("Notch Signaling Pathway" en hígado con EST) y las relaciones con otras proteínas externas a la ruta en un mismo tejido (Figura 13).

3.2.1.2.4. Ejemplo de una ruta biológica en cáncer

Al transferir la base de datos KEGG a MySQL se generó una tabla relacional conteniendo todas aquellas rutas relacionadas con cáncer denominada *Path2enet_CancerPathway*. Esta información puede ser analizada igualmente por la función en R *path2enet* para generar grafos de interacción de proteínas, locales y globales, de igual forma que se ha explicado anteriormente. Sin embargo, es una red grande y en estas ocasiones puede ser más productivo analizar la tabla de parámetros (tabla 15). Se puede observar como para esta red genes como AKT1/2/3 o EGFR juegan un papel clave, con alto número de interacciones, si bien ya está demostrado que defectos en el funcionamiento de estas proteínas están relacionados con varios tipos de cáncer. Un análisis completo de esta tabla de parámetros proporcionaría gran cantidad de información relevante, de forma sencilla y accesible. Además, en la figura 11 se muestra una visión global de este *Path2enet_CancerPathway*.

3.2.1.2.5. Combinando análisis estadístico y set de datos de expresión génica

Usando datos de ESTs, *Barcode*, chips de ADN o secuenciación de ARN, *Path2enet* genera redes tejido-específicas. Para ello, el usuario puede establecer filtros de expresión en función de los *vertex attributes* mencionados anteriormente. Un valor superior del umbral se considerará ON e inferior se considerará OFF. Una vez dicho umbral está definido, la función *graphParameters* en R analiza las características de la red tras seleccionar los nodos significativamente ON bajo esta especificación.

3.2.2. Caso de estudio: linfocitos B y linfocitos T

3.2.2.1. Introducción de set de datos propios

Path2enet permite introducir set de datos propios de origen transcriptómico. En este caso de estudio, vamos a analizar dos sets de datos de linfocitos B (CD19+) y linfocitos T (CD4+ y CD8+) y la ruta biológica “Notch Signaling Pathway”. Los datos provienen de cuatro sets de chips de ADN de Affymetrix de acceso público (descrito en Materiales y Métodos). La función expr2barcode realiza un análisis de los datos usando el algoritmo *Barcode* y determina que genes están ON ó OFF. La herramienta utiliza la librería *BrainArray* para eliminar aquellos probesets de Affymetrix ambiguos, así mejoramos los resultados de *Barcode*. Esta función genera un objeto con cuatro listas: (i) con el network creado en formato *igraph*, (ii) los resultados del análisis de *Barcode* sobre cada microarray, (iii) la media de valores de expresión por cada condición o tejido, y (iv) un objeto R tipo *ExpressionSet* con nuestros datos de expresión mapeados a IDs de UniProt.

3.2.2.2. Visualización en R

Con las funciones *graphTKplotterPATHW* y *graphTKplotterTissue* podemos visualizar estos datos de igual forma que hemos mencionado anteriormente (Figura 13).

3.2.2.3. Definiendo el estatus ON/OFF de los nodos de una red

Path2enet, como ya hemos mencionado, se sirve de *Barcode* para determinar qué gen/proteína está ON ó OFF. La tabla 16 constituye un buen ejemplo de ello, sobre los datos de estos tres

tipos celulares de linfocitos. En la red de linfocitos B se expresan 34/38 genes, mientras que en la red de linfocitos T son 22/24 los genes expresados.

3.3. Análisis proteogenómico cualitativo de la línea celular *Ramos* de linfoma de células B

En esta Tesis Doctoral se presentan dos estudios proteogenómicos, centrados en el análisis cualitativo de datos protéomicos y transcriptómicos de la línea celular Ramos de linfoma de células B.

3.3.1. Visión global de los datos proéomicos y transcriptómicos

Este trabajo se ha hecho en colaboración con la Unidad de Proteómica del Centro de Investigación del Cáncer (CiC-IBMCC, de USAL/CSIC, Salamanca), con los investigadores Paula Díez y Manuel Fuentes. Su grupo se encargó del diseño experimental del estudio, contribuyó a la generación de los datos proteómicos y se involucró en la discusión de resultados y la escritura de la publicación final "**Integration of Proteomics and Transcriptomics Datasets for the Analysis of a Lymphoma B-Cell Line in the Context of the Chromosome-Centric Human Proteome Project**" (Paula Díez, Conrad Droste et. al, 2015). En mi caso, me encargué del procesado de datos, análisis y creación de figuras para visualizar los resultados proteogenómicos. La figura 14 ilustra el modelo conceptual de integración de datos.

3.3.1.1. Procesando los sets de datos

3.3.1.1.1. Datos proteómicos

Dado que es un estudio cualitativo, los principales objetivos son (i) determinar si al menos un péptido es detectado para cada proteína en los datos proteómicos; (ii) establecer si una proteína es detectada por más de un péptido es signo de reproducibilidad, y (iii) en qué compartimentos subcelulares se encuentran las proteínas detectadas. No es un estudio centrado en isoformas, así que se eliminaron del set de datos.

Los pasos seguidos fueron: (1) eliminar la información sobre isoformas, (2) unificar las proteínas de diferentes compartimentos subcelulares de cada réplica biológica, (3) comparar los resultados de las tres réplicas biológicas en términos de detección de péptidos de una misma proteína, y (4) crear set de datos con diferentes niveles de confianza en la detección.

Así, se crearon tres sets de datos con niveles de confianza diferentes. *Intersection* hace referencia a aquellas proteínas con al menos dos péptidos detectados en las tres réplicas; alta confianza, pero baja cobertura de los datos. *Union* hace referencia a proteínas con al menos dos péptidos en cualquier réplica. *Maximum* referencia a proteínas con al menos un péptido en cualquier réplica.

3.3.1.1.2. Datos transcriptómicos

Este set de datos transcriptómico compila datos de expresión génica de acceso público, obtenidos de la base de datos Gene Expression Omnibus (GEO) con los siguientes IDs: GSE40168: GSM987747, GSM987748; GSM987749; plataforma GPL6244 (Affymetrix HuGene 1.0st). Debido a la plataforma de estos chips de ADN, no es posible usar *Barcode* para determinar si los genes están ON ó OFF, por lo que, tras el proceso de normalización de los datos, tomamos como genes

ON aquellos cuya expresión media a lo largo de los tres chips de ADN era superior al cuartil 25 de los datos.

3.3.1.2. Integración y comparativa de datos transcriptómicos y proteómicos

Para integrar estos dos sets de datos, en primer lugar, mapeamos ambos al identificador único de Ensembl. Nuestros tres sets de datos proteómicos *Intersection*, *Union*, y *Maximum* mapearon a 3433, 5540 y 8976 IDs respectivamente, y a 4088, 6175 y 9494 IDs de *Affymetrix* (Tabla 18). Esta integración hace posible comparar los niveles de expresión de ARNm correspondientes a las proteínas detectadas por algunos de sus péptidos (Figura 15), con mayor expresión génica conforme más confianza tenemos en el conjunto de péptidos/proteínas detectadas.

3.3.1.3. Integración y comparativa de datos transcriptómicos y proteómicos

Este estudio incluye un análisis de enriquecimiento funcional de aquellas proteínas encontradas en el set de mayor confianza *Intersection*, las 517 proteínas únicamente detectadas por proteómica, y los 1290 genes únicamente detectados por chip de ADN.

Al aplicar herramientas de enriquecimiento como *DAVID* o *GeneTermLinker* sobre el set *Intersection*, obtenemos que la lista está enriquecida, según lo esperado, en funciones de mantenimiento celular (housekeeping), regulación de moléculas biológicas importantes como ADN o ARN, y procesos de viabilidad celular como ciclo celular o crecimiento. Si ahora trabajamos con aquellas exclusivas del set proteómico, la ganancia se produce en términos relacionados con ADN mitocondrial, organelas ribosomales y citoplasmáticas. Es un resultado también esperado debido a que los chips de ADN ya que estos no incluyen sondas para ADN mitocondrial, y que proteínas relacionadas con el complejo mayor de histocompatibilidad siempre han sido elementos conflictivos a la hora de generar sondas para ellos debido a su variabilidad. Por último, el análisis sobre los genes exclusivos del set transcriptómico corresponde a proteínas de unión a ADN o nucleares, así como ARN no codificante. Ninguna fracción celular fue incluida en el set proteómico, así como los ARN no codificantes no pueden ser detectados, de ahí la exclusividad en el set transcriptómico.

3.3.1.4. Identificación de proteínas “perdidas” en el set proteómico

El término proteína “perdida” hace referencia a aquellos genes que muestran evidencias a nivel genómico y transcriptómico, pero no podemos detectar a nivel proteómico. La base de datos neXtProt recoge un set de datos de proteínas bajo esta denominación, de las cuáles podemos encontrar 370 dentro de nuestro set *Maximum*, 32 proteínas en *Union* y 4 en *Intersection*. Esta base de datos clasifica en cuatro categorías las proteínas, en función de si son proteínas detectadas a nivel transcriptómico, inferidas por homología, predichas o desconocidas (de PE2 a PE5), correspondiendo el nivel más bajo (PE1) a las proteínas detectadas a nivel proteómico. Así, la tabla 22 muestra la intersección según estos niveles con nuestros tres sets de datos.

3.3.2. Proteínas comunes al set transcriptómico y proteómico

Este estudio fue resultado de la colaboración con los investigadores Paula Díez y Manuel Fuentes, del Centro de Investigación del Cáncer (CiC-IBMCC, de USAL/CSIC, Salamanca). Los términos de colaboración fueron los mismos que en la descrita anteriormente, y dio lugar a una publicación bajo el título “**Comprehensive combination of affinity proteomics, MS/MS and**

RNA-Sequencing datasets for the analysis of a lymphoma B-cell line in the context of the Chromosome-Centric Human Proteome Project" (Paula Díez, Conrad Droste et. al, 2016). En este estudio, nos centramos en el conjunto de genes/proteínas comunes en los sets de datos proteómicos y transcriptómicos (Figura 18).

3.3.2.1. Procesando los sets de datos

Aquí trabajamos con tres sets de datos: secuenciación de ARN de la línea celular Ramos, set de datos proteómico de LC-MS/MS de la misma línea, y un set de 417 proteínas relevantes en el linfoma de Burkitt obtenido por afinidad a anticuerpos, con una técnica denominada SEC-MAP (Figura 17).

3.3.2.1.1. Datos proteómicos

Para el experimento con LC-MS/MS, se utilizaron 3 réplicas y 4 compartimentos subcelulares como fuente biológica. Las proteínas obtenidas de todas ellas ascienden a 5672, y el conjunto de las mismas fue denominado *complete mapping* de 7 subsets por el número de réplicas y compartimentos. Los genes asociados a dichas proteínas, así como su localización en el genoma humano, se obtuvo de la librería *biomaRt* en R.

Para el experimento con SEC-MAP, se usaron 549 anticuerpos con afinidad por 417 proteínas. Determinamos qué proteínas habían sido detectadas mediante un Qualitative Antibody Score (QAS) definido por el propio detector de señal de intensidad. De esta forma, un experto de la Unidad de Proteómica pudo determinar qué proteínas estaban siendo realmente detectadas por los anticuerpos al nivel más objetivo. Posteriormente, aquellas positivas fueron mapeadas a IDs de neXtProt.

3.3.2.1.2. Datos transcriptómicos

Los datos transcriptómicos corresponden a secuenciación de ARN de la línea celular Ramos (experimento SRX105534: http://www.ncbi.nlm.nih.gov/sra/SRX105534). Estos datos fueron normalizados a FPKM (Fragment Per Kilobase Of Exon Per Million Fragments Mapped), para obtener medidas de 20533 genes mapeados a IDs de Ensembl originalmente. Posteriormente, fueron traducidos a IDs de neXtProt (19518 genes), y establecido un umbral de FPKM > 1 para definir que el gen estaba expresado (9523 genes).

3.3.2.2. Integración y comparación de los datos

Puesto que los tres sets de datos contienen IDs de neXtProt, pudimos combinar los mismo con mayor facilidad. La tabla 22 contiene una pequeña comparativa del mapeo entre IDs de los tres sets de datos.

Los datos procedentes de MS/MS y secuenciación de ARN mostraron una correlación del 91%, con gran solapamiento a nivel analítico. Además, se observó una notable relación entre estos dos sets y el set de SEC-MAP en lo que a intensidad de señal transcriptómica se refiere (Figura 18), con curvas de densidad parecidas. Como era de esperar, la mayor parte de proteínas detectadas por MS/MS o SEC-MAP correspondían con niveles de FPKM superiores a 1, o a QAS mayor que 1. La correspondencia de valores se muestra en la Figura 19.

Una vez integramos los datos, el análisis revela un 56% de solapamiento (231/413) entre las tres aproximaciones. Específicamente, MS/MS un 65.6%, secuenciación de RNA un 80.6%, y SEC-MAP un 91.0% de las proteínas a detectar inicialmente. De todas ellas, hay 8 proteínas que ninguna de las tres técnicas ha conseguido detectar: JUN, CD44, CALB2, IL3RA, CTBP2, SEPT5, CDC14A, y MAPRE3 (Fig 20). El solapamiento entre técnicas también fue alto, como se muestra en la Figura 20.

3.3.2.3. Enriquecimiento funcional de la lista de proteínas

Para identificar las funciones principales de la lista de proteínas obtenidas únicamente por el solapamiento entre MS/MS y secuenciación de RNA, y aquellas que sólo obtiene SEC-MAP, utilizamos herramientas como *DAVID* o *GeneTermLinker*.

Para las 57 proteínas obtenidas exclusivamente por SEC-MAP, funciones como proteínas de membrana, adhesión celular, o proteínas transmembrana revelan que la inmunidad de anticuerpos mejora la obtención de proteínas de membrana, difícilmente detectadas por otras técnicas. Por otro lado, para las 29 proteínas exclusivas de MS/MS y secuenciación de ARN, y por tanto con valores de QAS menores de 1, revelan funciones de proteína nuclear, proteínas señalizadores fosforiladas y regulación GTPasa. Estos resultados pueden esperarse ya que proteínas con alto nivel de modificaciones postraduccionales requieren de anticuerpos muy específicos, y, además, los protocolos de aislamiento han de ser más agresivos para estos compartimentos celulares.

3.3.2.4. Proteínas “perdidas”

De forma análoga al apartado 3.3.1.4, en la tabla 23 se muestran datos acerca de los niveles de proteínas “perdidas” obtenidas mediante estas técnicas proteómicas en el set complete mapping definido anteriormente.

3.4. Proteoma y fosfoproteoma en la Leucemia Linfocítica Crónica de células B

En este apartado nos preparamos a analizar el proteoma y fosfoproteoma de pacientes con B-CLL y CLL-tipo linfocitosis monoclonal de células B (MBL), mediante el uso de un espectrómetro de masas de alta resolución. De igual forma, este estudio fue resultado de la colaboración con los investigadores Paula Díez y Manuel Fuentes, del Centro de Investigación del Cáncer (CiC-IBMCC, de USAL/CSIC, Salamanca). Los términos de colaboración fueron los mismos que en la descrita anteriormente, y dio lugar a una publicación bajo el título “**Revealing Cell Signaling Pathways in ChronicLymphocytic Leukemia Tumor B-cells by Integration of Global Proteome and Phosphoproteome Profiles**” (Paula Díez, Conrad Droste et. Al, en revisión 2017).

3.4.1. Procesado de los datos

A diferencia del anterior, este estudio es cuantitativo además de cualitativo, y está centrado en la cantidad de proteína detectada. Partimos de 4 muestras de pacientes con CLL, y uno con CLL-tipo MBL (Figura 21). En total, pudimos detectar 13504 péptidos inequívocos correspondientes a 2970 proteínas en el set proteómico, y 594 péptidos inequívocos correspondientes a 327 proteínas en el set fosfoproteómico. Para cada péptido, se calculó la media de las dos réplicas

existentes por paciente, y si dicha media estaba por debajo del umbral de 50000 en lo referente a señal, se definía el péptido como no detectado (corregido a 0). Además, información sobre la proteína (de UniProt), su nivel de proteína "perdida" y su localización en el genoma, fue añadido para análisis posteriores.

3.4.2. Análisis cualitativo

El análisis cualitativo en este caso no requiere de mayor profundidad, puesto que fue determinado el umbral en 500.000 de valor, para decidir si un péptido estaba expresado o no. Así, se observó que sólo el 17% del fosfoproteoma se encontraba presente en los cinco pacientes. Proteínas involucradas en la fosforilación de proteínas, en la señalización celular, y en el transporte intracelular de proteínas. Además, ciertos pacientes mostraban mayor concentración de proteínas con la unión a ARN o el spliceosoma.

3.4.3. Análisis cuantitativo

Los principales obstáculos surgidos para el análisis cuantitativo fueron (i) el bajo número de pacientes y réplicas y (ii) que la concentración de las proteínas solo puede ser comparada de forma entre pacientes, y no de una proteína a otra. Así, utilizamos tanto el tanto por uno del valor respecto al máximo detectado por péptido, como el ranking de concentraciones (1 para el mayor, 5 para el menor) por péptido para trabajar en este apartado.

3.4.3.1. Expresión diferencial de péptidos/proteínas

Tras dar un ranking a los valores de los péptidos a lo largo de las muestras, observamos cómo el 63% de los péptidos con mayor valor procedían de la muestra A, mientras que el 70% de las concentraciones más bajas con la muestra E. En el resto también vimos esta predominancia de score para cada una de las muestras.

Basándonos en la posición en el ranking (1-5), pudimos hacer un clustering jerárquico de las 5 muestras para ver si existían relaciones entre ellas (Figura 27), dónde se podían observar dos grupos predominantes con muestras (i) A y C, y (ii) B, D, y E. Sin embargo, cuando atendíamos al clustering jerárquico del fosfoproteoma, la muestra C se intercambiaba de grupo con la muestra B. Por otro lado, si atendemos respecto al máximo valor por péptido, el clustering jerárquico obtenido para el proteoma separa a las muestras de igual forma.

3.4.3.2. Análisis y visualización de proteínas en las deleciones del cromosoma 11 y 13

De igual forma, usando los valores rateados al máximo valor de concentración por péptido, hemos mapeado las posiciones cromosómicas relevantes en CLL y CLL-tipo MBL para observar si existe alguna relación. De las proteínas aquí detectadas y localizadas en las regiones del11q22.3, del11q23.3, y del13q14, valores de expresión más bajos para ATM y CUL5 en la primera región, de MLL, SCN4B, CD3D, ARCN1 y TREH en la segunda región, así como RB1 en del13q14 fueron hallados para la muestra C (Figura 29).

4.DISCUSIÓN

4.1. Path2enet

4.1.1. Aspectos técnicos

El código escrito en R para *Path2enet* es funcional y compatible con versiones de R-2.6. Las funciones han sido escritas para ser independientes de algunas librerías pesadas de R, y su implementación para la creación y modificación de bases de datos mediante MySQL como reto ha resultado satisfactoria. En ese sentido, hemos podido probar correctamente: (i) la construcción de una base de datos KeggSQL desde archivos KGML; (ii) las conexiones entre R y MySQL; (iii) la creación de redes de proteínas combinando las bases de datos APID, KEGG y datos transcriptómicos de ESTs, chips de ADN o secuenciación de ARN; (iv) la visualización de redes proteicas con la librería *igraph*; y (v) el análisis topológico de redes usando esta librería *igraph*.

4.1.2. Selección de fuentes de datos

4.1.2.1. Sets de datos de interacción

La base de datos APID fue elegida por estar centrada en interacciones físicas proteína-proteína probadas experimentalmente de 448 organismos, a diferencia de otras bases de datos de interacción de genes/proteínas como STRING, GeneMANIA o ConsensusPaTHDB que incluyen interacciones predichas y otros tipos de interacciones no físicas.

Por otro lado, elegimos KEGG PATHWAY como base de datos de rutas biológicas por su calidad en la revisión de la información y su popularidad en la comunidad científica, incluyendo 517 pathways y 515499 referencias. Esta base de datos no es solo un compendio de archivos KGML, añade una serie de ventajas respecto a otras bases de datos: (i) un mapa de toda proteína o gene del ID de KEGG al ID de UniProt y (ii) incluye una estructura relacional SQL, ambos aspectos muy útiles y en concordancia con nuestro trabajo. Sin embargo, debido a algunos problemas técnicos, la creación de nuestra base de datos MySQL con otras bases de datos de rutas biológicas extendidas como Reactome o Pathway Commons sigue en desarrollo. El principal reto aquí es el mapeo a IDs de UniProt de forma correcta e inequívoca.

Estas dos bases de datos se complementan a la perfección, puesto que la presencia de una interacción entre dos proteínas en APID no descrita en KEGG nos indica que aún no está bien caracterizada o es reciente, mientras que de forma análoga podemos observar cuáles de estas son consideradas canónicas puesto que se incluyen en KEGG.

4.1.2.2. Set de datos de expresión génica

Path2enet integra datos transcriptómicos de tres plataformas diferentes (ESTs, chips de ADN y secuenciación de ARN) en las redes biológicas generadas, de cuatro fuentes distintas: UniGene, *Barcode*, Human Body Map 2.0 y Human Protein Atlas. De esta forma, se pretende valorar de forma eficiente si un gen está activo (ON/OFF) en un tejido o tipo celular específico, según la plataforma elegida (ESTs con al menos un transcrito, *Barcode* con su algoritmo o por encima del cuartil 25, y *RNA-Seq* con FPKM mayor que 1).

Cabe destacar, que ningún set de datos proteómico ha sido incluido en *Path2enet* hasta el momento, debido a la variabilidad de técnicas existentes y a la falta de algún protocolo estándar para una óptima detección de péptidos/proteínas en muestras biológicas.

4.1.3. Beneficios de Path2enet para la investigación biológica

Basándonos en las categorías IT estándares para el análisis básico de software, en la tabla 28 se incluye una breve enumeración de los beneficios y resultados que proporciona *Path2enet*, en función de la entrada, proceso y salida de cada una de las funciones o pasos de esta herramienta.

4.1.4. Características de Path2enet diferentes a otras librerías similares de R

Aquí describimos brevemente aquellas librerías de R que incluyen información sobre rutas biológicas o interacción de proteínas para sus análisis. Algunas de ellas también integran datos de expresión en sus redes.

La librería de R *KEGGgraph* usa los archivos KGML para crear grafos directos o indirectos, mapea los genes a IDs de Human Gene Symbols, y utiliza *igraph* como formato de objeto para crear las redes mediante la conjunción de rutas biológicas de KEGG. Esta librería es usada por otras como *Pathview* ó *MetaboSignal* para realizar análisis de rutas metabólicas. Igualmente, *rBiopaxParser* y *PaxtoolsR* trabajan sobre archivos tipo BioPAX (estándares)de KEGG para generar redes más complejas. Sin embargo, nuestro *Path2enet* no sólo genera redes más grandes por la conjunción de rutas biológicas, sino que también genera una base de datos relacional MySQL que permite acceder a la información de forma óptima y relacionarla con otras bases de datos directamente. Como punto negativo, *Path2enet* no está diseñado para acceder a archivos BioPAX, de uso estándar en rutas biológicas, ya que impediría la creación de una base de datos MySQL y las rutas deberían consultarse individualmente.

La integración de datos transcriptómicos de tres plataformas diferentes, crea un mapa robusto de expresión de genes tejido-específica que permite al usuario tener una idea global sobre el estado de rutas o redes biológicas en contextos biológicos concretos, e integrar datos de tipo proteómico. En este sentido, no existe ninguna otra librería en los repositorios de R con dicha información integrada.

Otro de los objetivos iniciales de *Path2enet* era definir el estatus ON/OFF de los genes/nodos en una red generada. Ninguna de las herramientas mencionadas anteriormente atiende a ello, si bien, si permiten incluir datos transcriptómicos orientados hacia la expresión diferencial (medida relativa entre condiciones). Herramientas como *SubpathwayMiner* y BioNet permiten encontrar subgrafos desregulados en base a expresión génica diferencial, pero con metodología estadísticamente diferente. Así, la implementación de *Barcode* y *BrainArray* ha permitido dicha aproximación, haciendo también posible eliminar aquellos nodos cuyo estatus OFF y que parecen indicar que no están siendo relevantes en la red biológica que se esté estudiando.

4.1.5. Caso de estudio Linfocitos-T y linfocitos-B

Hemos probado *Path2enet* sobre la ruta biológica de NOTCH “NOTCH Signaling Pathway” de KEGG, para demostrar que la integración de datos transcriptómicos y su estatus ON/OFF alteran la interpretación de esta ruta canónica de KEGG. Dichas modificaciones no son aleatorias, y responden a un significado biológico. Mostramos que dos de los cuatro parálogos existentes de NOTCH son expresados en células B y T (NOTCH1 en células B CD19+ y células T CD4+, mientras

que NOTCH2 en los tres tipos celulares) y que reguladores clave como DTX1 y HES1 están altamente expresados en células B.

4.2. Análisis proteogenómico cuantitativo de la línea celular Ramos

En esta sección vamos a detenernos en las dos publicaciones sobre la línea celular Ramos (Paula Díez, Conrad Droste et. al 2015) y (Paula Díez, Conrad Droste et. al, en proceso 2017).

4.2.1. Comparación de los sets de datos transcriptómicos y proteómicos

4.2.1.1. Identificación de genes/proteínas expresados

El set de datos proteómico de LC-MS/MS tenía cuatro fracciones subcelulares que fueron combinadas y comparadas para generar los subsets de proteínas *Intersection*, *Union* y *Maximum*, como ya explicamos anteriormente. Realizamos un análisis de enriquecimiento funcional sobre *Intersection* que nos permitió descubrir cómo funciones de mantenimiento celular, ciclo celular y crecimiento celular eran las más representadas, hecho esperable y que reafirmaba la fiabilidad de dicho subset. Por otro lado, al mapear las proteínas del subset *Maximum* a los cromosomas, observamos que el 30% de los genes/proteínas conocidos están presentes. Muchas de las proteínas del mismo estaban relacionadas o interactúan con MYC, un proto-oncogén con papel clave en el linfoma de Burkitt. En el segundo estudio, pese a la alta fiabilidad de los subset proteicos del primer análisis, decidimos ampliar el *Intersection* dataset bajando el umbral a al menos 1 péptido de la proteína por réplica (5 muestras con dos 2 réplicas).

Si nos detenemos en el set de datos proteómico de SEC-MAP, esta técnica permitió analizar un subset concreto de proteínas mediante la incorporación de 549 anticuerpos, correspondientes a 417 proteínas. Además, la técnica incorpora un score de calidad (QAS) por el cuál determinar aquellas proteínas que estaban siendo realmente detectadas, y es tremendamente útil a la hora de detectar proteínas de una forma objetiva (no centrada en sus péptidos).

Desde nuestro punto de vista, la implementación del algoritmo *Barcode* para determinar el status ON/OFF de cualquier gen en el set de datos transcriptómico (chip de ADN) fue un éxito para *Path2enet*. Sin embargo, la plataforma de Affymetrix utilizada aquí (Affymetrix Human Gene 1.0 ST) no era compatible con dicho algoritmo, por lo que optamos por establecer un umbral mínimo de expresión de cuantil 75. En comparación con otras publicaciones es un umbral alto, pero nos aseguramos no tener señal ruido en los resultados, aumentar la probabilidad de que dichos genes estén ON, y mantener un alto grado de confianza en los resultados (8976 Affy probesets). Para el set de datos de secuenciación de ARN seguimos una filosofía parecida, y tomamos como umbral FPKM ≥ 1. Este umbral está extendido en la comunidad científica y se considera fiable y reproducible entre experimentos. De esta forma, nos quedamos con 5157 genes que mapeaban a 5672 proteínas, un tamaño muy similar al obtenido mediante el proteoma de LS-MS/MS.

4.2.1.2. Integración de los sets de datos

La integración de datos incluye mapeo entre identificadores, aquí realizado en dirección IDs de neXtProt a IDs de Ensembl porque una proteína procede de un gen (en la mayoría de los casos), pero un gen puede dar lugar a varias proteínas. Del subset proteómico más amplio, *Maximum*, únicamente 516 proteínas no fueron detectadas por chip de ADN (solapamiento del 94%), y sólo

35 proteínas no fueron detectadas por secuenciación de ARN (99.8%), lo que nos da una idea del alto grado de reproducibilidad entre sets, la precisión de las técnicas y la profundidad de análisis de la integración de dichos métodos. Además, los niveles de expresión génica de aquellas proteínas detectadas eran significativamente mayores que del total de genes medidos, aumentando los mismos conforme crece la fiabilidad del subset proteómico (de *Maximum* a *Intersection*).

Además del alto grado de solapamiento, esta aproximación nos permitió analizar las proteínas que se escapaban de métodos de detección transcriptómicos (516 proteínas). El análisis de enriquecimiento funcional revelaba funciones relacionadas con la mitocondria y organelas ribosomales. Como ya mencionamos anteriormente, los chips de ADN no incluyen sondas para ADN mitocondrial, así como se evidenciaba la falta de detección de Inmunoglobulinas y MHC, debido a la variabilidad de estas proteínas. Si atendemos a los sets transcriptómicos, 1290 genes/proteínas eran exclusivos de estos métodos, y estaban asociados a la fracción nuclear de la célula y ARN no codificante. El aislamiento de la fracción nuclear no está optimizado en estos momentos para las técnicas ómicas descritas, y el ARN no codificante no es posible de detectar.

4.2.2. Aproximación global o pre-selección de proteínas con SEC-MAP

Si atendemos a proteínas específicas, el grado de solapamiento no es indicativo de calidad en la detección. En el apartado 3.3.2.2, se mencionaba como la técnica SEC-MAP era capaz de mapear 57/413 (13,8%) proteínas exclusivas, dando a entender que dicha técnica permite la detección de un significativo set de proteínas que de otra forma no podríamos detectar. Por supuesto, dicho resultado necesita de una validación dada la falta de confirmación por técnicas como la secuenciación de ARN, pero su valor pone de manifiesto la necesidad de desarrollar dicha técnica. Además, el análisis de enriquecimiento de este subset de 57 proteínas, revelaba proteínas relacionadas con diferentes membranas celulares o la matriz extracelular. La detección con anticuerpos de estas proteínas parece resultar más eficiente, así como el hecho de que proteínas complejas con alto número de modificaciones postraduccionales no aparezcan en esta fracción de SEC-MAP era esperado. Sin embargo, SEC-MAP está limitado por la calidad de los anticuerpos y por el alto coste y tiempo que consume el diseño experimental.

En resumen, se trata de una técnica prometedora y de fácil integración con otras técnicas masivas tanto transcriptómicas como proteómicas, que permite el estudio de un set más pequeño de proteínas de nuestro interés.

4.2.3. Beneficios del análisis proteogenómico

La integración de metodologías complementarias permite mejorar la caracterización de tipos celulares, tejidos o enfermedades, puesto proporciona información a varios niveles de regulación biológica. En esta Tesis Doctoral hemos caracterizado de esta forma la línea celular Ramos del linfoma de Burkitt, así como hemos comparado la capacidad de detección y el solapamiento de genes/proteínas detectadas por estos métodos, dando una visión global del estado del arte de dichas técnicas.

4.3. Análisis proteómico de células B de pacientes con CLL ó MBL

4.3.1. Datos proteómicos

Tras la eliminación de péptidos ambiguos y valores por debajo del umbral de detección (valor de 500000), obtuvimos 2970 proteínas mapeadas por 13504 péptidos. Para el fosfoproteoma, sólo 327 proteínas mapeadas por 594 péptidos. Como era de esperar, el fosfoproteoma es mucho más pequeño que el proteoma, incluyendo además 253 proteínas detectadas por primera vez, puesto que ningún experimento LC-MS/MS previo había podido determinar su presencia en muestras biológicas.

4.3.2. Comparación en pacientes con cáncer

El fosfoproteoma obtenido presenta una reproducibilidad muy baja entre los cinco pacientes (17% de solapamiento), mientras que el proteoma sí es fiable (73% de solapamiento). De hecho, aunque la mayoría de proteínas eran muestra-específicas, las proteínas relacionadas con receptores de citoquinas y TLR están infra-representadas en comparación con la ruta de señalización de BCR en todas las muestras analizadas. Además, muchas proteínas relacionadas con esta ruta de señalización están presentes en todas las muestras, sugiriendo que esta ruta tiene un papel clave en el mantenimiento de la supervivencia de estas células.

Debido a la imposibilidad de comparar las concentraciones de péptidos entre ellos (únicamente entre muestras), no se pudo realizar un análisis de expresión de proteínas común. Se adoptaron dos medidas para intentar solventar este problema, como fueron (i) el ranking de cada péptido en las 5 muestras y (ii) relativizar con el porcentaje de señal respecto a la máxima señal de cada péptido. Ambas medidas nos permitieron realizar un clustering jerárquico, que relevó dos grupos (A y C, y B, D, E con ranking para el proteoma) como representación de las tendencias generales dentro del set de muestras.

En general, pese a obtener resultados significativos tras el análisis de enriquecimiento del proteoma y del fosfoproteoma, fue que esta aproximación debería ser mejorada en lo referente al (i) tamaño muestral, (ii) valores de referencia con los que estandarizar los datos y (iii) parámetros experimentales que permiten eliminar péptidos con señal de ruido de forma más eficiente.

5.CONCLUSIONES

El trabajo presentado en esta Tesis Doctoral se ha centrado en el desarrollo y aplicación de algoritmos bioinformáticos para integrar, analizar y visualizar varias fuentes de datos transcriptómicos y proteómicos de humano. Después de todo este trabajo, hemos llegado a las siguientes conclusiones:

1.- Hemos diseñado y creado una aplicación bioinformática gratuita llamada *Path2enet*, que permite la integración de información sobre rutas biológicas, proteínas y expresión génica humana en un contexto relacional gracias a la transferencia de dicha información a una red biológica. La generación de redes biomoleculares permite explorar las interacciones y conexiones entre proteínas, a la vez que su centralidad en la red, todo ello hecho en un contexto tejido o tipo celular específico graticas a la integración de datos de expresión génica sobre qué genes o proteínas están ON en un determinado contexto biológico.

2.- Hemos diseñado y aplicado varias estrategias bioinformáticas para analizar de forma integrativa datos proteómicos y transcriptómicos. Esta integración y análisis comparativo ha sido aplicado a tres casos específicos de estudio de datos humanos (linfocitos B del linfoma de Burkitt y pacientes de leucemia) mostrando un alto grado de reproducibilidad en el perfil global de genes y proteínas detectados por ambos métodos ómicos.

3.- El análisis integrativo de datos proteómicos y transcriptómicos muestra que, a parte de un alto solapamiento entre técnicas, la sensibilidad y especificidad de estas tecnologías no es la misma. En este sentido, siempre hemos observado una alta cobertura del genoma y muy buena reproducibilidad con técnicas transcriptómicas. Sin embargo, las técnicas proteómicas, y en especial las fosfoproteómicas, muestran mayor sensibilidad para detectar proteínas que no podemos detectar mediante técnicas transcriptómicas, generando un marco de trabajo más propicio a nuestro parecer para la identificación de biomarcadores específicos.

4.- Nuestro trabajo demuestra que la aplicación de estudios proteogenómicos con estrategias bioinformáticas robustas proporcionan un excelente marco de investigación para acometer el estudio biomolecular integrativo de varias capas de información ómica. Esta estrategia es esencial para el estudio de enfermedades complejas como el cáncer, y debería ser una práctica común en colaboraciones e investigación multidisciplinar dónde la experiencia bioinformática es crítica.