UNIVERSIDAD NACIONAL DEL SUR

TESIS DE DOCTOR EN CIENCIAS DE LA COMPUTACIÓN

DESARROLLO DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO Y COMPUTACIÓN EVOLUTIVA MULTIOBJETIVO PARA LA INFERENCIA DE REDES DE ASOCIACIÓN ENTRE VÍAS BIOLÓGICAS

Julieta Sol Dussaut
2016

Prefacio

Esta Tesis se presenta como parte de los requisitos para optar al grado Académico deDoctor en Ciencias de la Computación, de la Universidad Nacional del Sur y no ha sidopresentada previamente para la obtención de otro título en esta Universidad u otra. Lamisma contiene los resultados obtenidos en investigaciones llevadas a cabo en el ámbitodel Departamento de Ciencias e Ingeniería de la Computación durante el períodocomprendido entre el 30 de Agosto de 2011 y el 10 de Febrero de 2016, bajo la direcciónde Dr. Ignacio Ponzoni, profesor del Departamento de Ciencias e Ingeniería de laComputación e investigador independiente del CONICET, y de la Dra. Jessica A. Carballido,profesora del Departamento de Ciencias e Ingeniería de la Computación e investigadoraadjunta del CONICET.

Julieta Sol Dussaut

Bahía Blanca, 10 de Febrero 2016

Universidad Nacional del Sur
Secretaría General de Posgrado y Educación Continua

La presente tesis ha sido aprobada el .../.../....., mereciendo la calificación de ..... (.....)

Agradecimientos

Esta tesis se llevó a cabo gracias a varias personas. En primer lugar quisiera agradecer al Dr. Ignacio Ponzoni y la Dra. Jessica Carballido quienes me iniciaron e hicieron crecer en el área de investigación. Agradezco también al grupo de trabajo, quienes en conjunto con mis directores me brindaron su ejemplo y paciencia, entre ellos el Dr. Cristian Gallo y la Dra. Rocío Cecchini.

Además me gustaría agradecer a mis compañeros de trabajo y amigos por brindarme su aliento y cariño, especialmente a la Ing. Jimena Martínez quien me brindó su apoyo y amistad a lo largo de la carrera de grado y posgrado.

Por supuesto agradezco a mi familia quienes me han alentado a seguir este camino de investigación, sobre todo a mi mamá y a mi abuela, quienes promovieron este aspecto de mi vida. También agradezco profundamente a Carlos quien me acompañó brindándome su amor y apoyo incondicional en estos años.

Por último quiero agradecer al Consejo Nacional de Investigaciones Científicas y Técnicas y al Departamento de Ciencias e Ingeniería de la Computación por haberme facilitado los medios para hacer este trabajo posible.

Resumen

En la biología de sistemas, una ruta biológica representa una secuencia de reacciones o interacciones entre un grupo de genes expresados que participan en un proceso biológico. Durante la última década, el análisis de las rutas biológicas se ha convertido en una estrategia clave para la comprensión de los significados biológicos de experimentos de alto rendimiento sobre un grupo de genes. Detrás de la idea del análisis de estas rutas existe el supuesto de que, para muchos fenómenos celulares complejos, resulta muy difícil encontrar una explicación mediante estudios que sólo se centran en una mirada al nivel de los genes.

En particular esta tesis se centra en la investigación de técnicas de análisis de diafonía (*cross-talk*) entre rutas biológicas (*pathways*), enriqueciendo esta información por datos de experimentos de *microarray* mediante *biclustering*. De esta forma, se busca proveer una metodología bioinformática que identifique relaciones entre rutas biológicas y las explique, proporcionando información útil para asistir a expertos en biología molecular.

Para cumplir este objetivo se desarrollaron métodos computacionales para el análisis tanto topológico como de enriquecimiento a nivel de rutas biológicas. Una de las herramientas desarrolladas, BAT(Gallo, Dussaut, Carballido, & Ponzoni, 2010), plantea la ejecución del algoritmo BiHEA(Gallo, Carballido, & Ponzoni, 2009), que realiza biclustering sobre los datos. Esto permite la identificación de grupos de genes co-expresados bajo ciertos subconjuntos de condiciones experimentales. Esta herramienta es utilizada en conjunto con otra, denominada PET, diseñada para utilizar datos topológicos relevantes a nivel de genes y proyectarlos a nivel de rutas biológicas para una mejor comprensión de los mecanismos de señalización que coordinan distintos procesos celulares.

Se estudiaron y validaron estos métodos con datos de la enfermedad de Alzheimer, contrastando los resultados con los obtenidos por otros métodos publicados recientemente. De este modo, se puso en evidencia la relevancia de combinar técnicas de análisis topológico con enriquecimiento basado en datos de expresión y detección de sincronización entre rutas biológicas mediante el uso de métodos de *biclustering* como una estrategia integral para la identificación de diafonía entre procesos biológicos.

Abstract

In systems biology, a pathway represents a sequence of reactions or interactions between a group of expressed genes involved in a biological process. During the last decade, the analysis of biological pathways has become a key strategy for the understanding of biological meanings in high throughput experiments on a group of genes. Behind the idea of the analysis of these pathways there is the assumption that, for many complex cellular phenomena, it is very difficult to find an explanation through studies that focus only at a gene level.

In particular, this thesis focuses on the investigation of cross-talk analysis techniques between biological pathways, also enriching this information by *microarray* experiments data using*biclustering*. By means of this combination, the idea is to count with a bioinformatics approach that identifies and explains relationships between biological pathways thus providing useful information to assist experts in molecular biology information.

To meet this objective, computational methods for analysis of biological pathways, including enrichment analysis, and analysis at a topological level,has been developed. One of the tools developed, BAT (Gallo, Dussaut, Carballido, & Ponzoni, 2010)raises the algorithm execution BiHEA (Gallo, Carballido, & Ponzoni, 2009), which is a biclustering multi-objective algorithm. This allows the identification of clusters of co-expressed subsets of genes under certain experimental conditions. This tool is used in conjunction with other, called PET, designed to use topological data relevant at gene level and project biological pathways for better understanding of the signaling mechanisms that coordinate various cellular processes.

We studied these methods and validated them with data from Alzheimer's disease, contrasting results with those of other recently published methods. Thus, is highlighted the importance of combining topological analysis techniques with enrichment expression data based on detection and synchronization between biological pathways using methods of biclustering as a comprehensive strategy for identifying crosstalk between biological processes.

Índice

CAPÍTULO 1: Introducción

En esta era en la que la salud es una de las industrias más grandes y crecientes, hay un gran interés en entender qué ocurre con nuestras células y órganos a nivel molecular. Afortunadamente, las mejoras e innovaciones en la tecnología continúan estimulando la calidad y los tipos de datos biológicos que pueden ser obtenidos a nivel de genoma completo. De esta forma, una gran cantidad de información de varios años de investigación detallada puede ser encontrada en formas de anotaciones o bases de datos computacionales. Dicha información organizada en conjuntos de datos apropiadamente combinados, tiene el gran potencial para posibilitar descubrimientos novedosos que lleven a avances en biología y medicina.

El presente trabajo de tesis se enfoca en un problema de especial interés dentro del análisis de datos biológicos en el área de Bioinformática. En líneas generales, el objetivo es estudiar redes biológicas para entender el significado de un conjunto de datos de genes derivado de experimentos de alto rendimiento (*high-thoughput*), tal como los experimentos de microarrays.

En este sentido todos los organismos vivos pueden ser considerados como redes complejas de biomoléculas y reacciones bioquímicas. Las mismas representan la suma de interacciones fuertemente controladas entre dichos componentes y determinan la forma y función de un organismo. En el estudio de redes biológicas, las series de acciones entre biomoléculas que llevan a efectos biológicos específicos son comúnmente descritas como

rutas biológicas (*pathways*). En otras palabras, una ruta biológica es un subconjunto significativo de biomoléculas o genes y reacciones cuya interacción cumple una función específica en una célula u organismo. Algunas redes biológicas describen procesos metabólicos, mientras que otras están involucradas en enfermedades.

Numerosos conjuntos de datos genéticos y genómicos relacionados con enfermedades han estado disponibles durante las últimas décadas. Es un gran desafío evaluar hoy en día estos conjuntos de datos heterogéneos para priorizar genes relacionados con enfermedades y realizar su correspondiente análisis y validación.

De esta forma el principal foco de esta tesis consiste específicamente en la investigación de técnicas de computación evolutiva, minería de texto y análisis estadístico para la inferencia de redes de rutas biológicas a partir de datos de expresión de microarray en conjunto con bases de datos conocidas para la utilización de topología de rutas biológicas con datos curados.

Esta tesis se organiza como sigue: los capítulos 2 y 3 introducen conceptos básicos de biología molecular y tecnología de microarrays y están orientados a introducir al lector ajeno a estos conocimientos con los términos que serán utilizados en el resto de la tesis. El capítulo 4 introduce la problemática de biclustering, presenta a la computación evolutiva y proporciona una revisión del estado del arte de las técnicas disponibles en la literatura para biclustering. En el capítulo 5 se introducen conceptos básicos de minería de datos y su utilidad a la problemática establecida. En el capítulo 6 y 7 se presentan las contribuciones de esta tesis, relacionadas con la inferencia de redes de rutas biológicas. Finalmente, en el capítulo 8 se presentan las conclusiones generales obtenidas a partir de todo el trabajo involucrado con las investigaciones de esta tesis.

CAPÍTULO 2: Conceptos básicos de biología molecular

Este capítulo está destinado a aquellos lectores que no poseen los conocimientos básicos de biología molecular. Es una presentación muy general de conceptos como célula, ADN, ARN, genes y vías biológicas, necesarios para comprender los capítulos subsiguientes.

2.1 Células

La célula es el elemento de menor tamaño que puede considerarse vivo (Alberts, Johnson, Lewis, Raff, & Roberts, 2004). La teoría celular fue primeramente formalizada en 1838 para los vegetales y en 1839 para los animales por Theodor Schwann y Matthias Schleiden (Aréchiga, 1996), y postula que todos los organismos están compuestos por células y que todas las células derivan de otras precedentes. Esta teoría se puede resumir en tres principios:

1. Todos los seres vivos están compuestos por células o por sus productos de secreción.
2. En una célula caben todas las funciones vitales, de manera que basta una célula para tener un ser vivo.
3. Todas las células proceden de células preexistentes (Standafer & Wahlgren, 2002). Es la unidad de origen de todos los seres vivos.

Existen dos clases de células: procariotas y eucariotas (figura 1), dependiendo de su estructura interna. Las células procariotas poseen estructuras más simples, representadas por las bacterias y cianobacterias, mientras que las células más complejas, llamadas eucariotas, se encuentran en todos los otros tipos de organismos.

Sin embargo, pese a sus diferencias de complejidad, ambas clases de células poseen una región nuclear que contiene el material genético. El material genético de una célula procariota está presente en un nucleoide, que es una región de la célula la cual no contiene ninguna membrana que la separe del citoplasma que la rodea. En contraste, en las células eucariotas el material genético se encuentra en un núcleo, que consiste en una región delimitada por una estructura membranosa compleja llamada envoltura nuclear. Esta diferencia en la estructura nuclear es la base para los términos procariota (pro=antes, cariota=núcleo) y eucariota (eu=verdadero, cariota=núcleo)(Karp, 2009).

En ambas clases de células se presentan moléculas importantes en la biología, las **proteínas** y los **ácidos nucleicos (ADN y ARN)**.

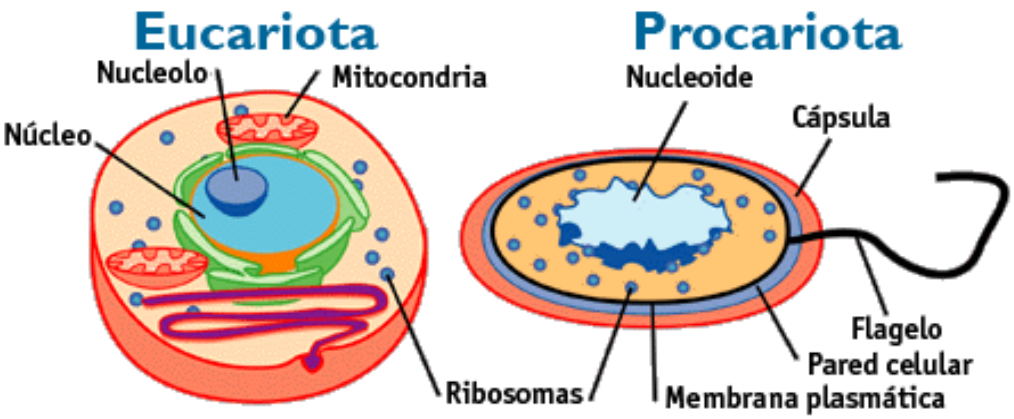

Figura 1 – Diferencias entre una célula Eucariota y una Procariota.

2.2 Proteínas

La mayoría de lo que nos rodea está formada por grupos de átomos unidos que forman conjuntos llamados moléculas. En particular, las proteínas ocupan un lugar de enorme importancia entre las moléculas constituyentes de los seres vivos. Son largas cadenas de aminoácidos unidas por enlaces peptídicos entre el grupo carboxilo (-COOH) y el grupo amino ($-NH_2$) de residuos de aminoácido adyacentes. La secuencia de aminoácidos en una proteína está codificada en su *gen* (una porción de ADN) mediante el código genético. Esta secuencia de aminoácidos es conocida como *estructura primaria* de la proteína, dentro de una jerarquía estructural de 4 partes (figura 2). La *estructura secundaria* es la disposición espacial local del esqueleto proteico, gracias a la formación de puentes de hidrógeno entre los átomos que forman el enlace peptídico. Luego, la *estructura terciaria* es el modo en que la cadena polipeptídica se pliega en el espacio. Por último, la *estructura cuaternaria* de una proteína existe si intervienen más de un polipéptido.

Las proteínas son imprescindibles para el crecimiento del organismo y realizan una enorme cantidad de funciones diferentes. Prácticamente todos los procesos biológicos dependen de la presencia o la actividad de este tipo de moléculas. Las *proteínas estructurales* forman parte de la estructura celular, las *enzimas* catalizan casi todas las reacciones bioquímicas que ocurren en la célula, las *proteínas regulatorias* controlan la expresión de los genes o la actividad de otras proteínas, y las proteínas de transporte llevan otras moléculas a través de la membrana celular o alrededor del cuerpo(Amaratunga & Cabrera, 2009).

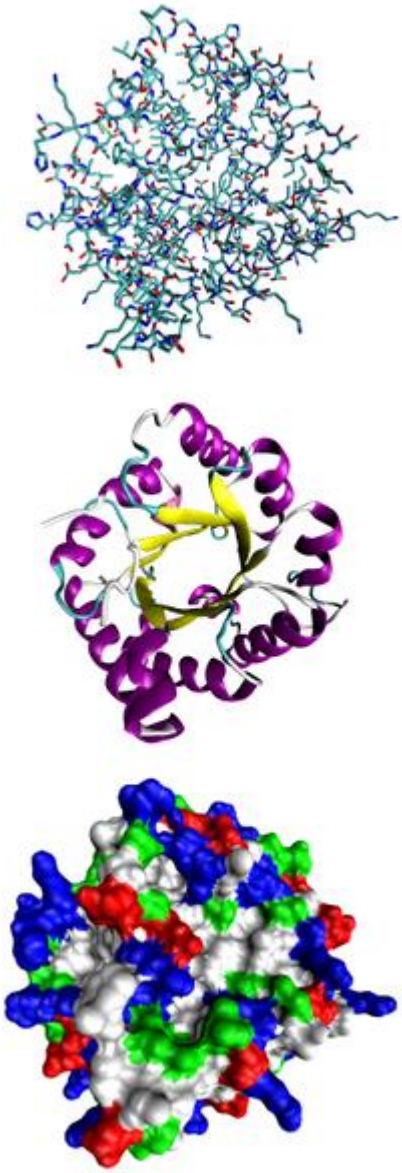

Figura 2 - Representación de la estructura proteica a tres niveles: Primario, Secundario y Terciario de arriba hacia abajo.

2.3 ADN (Ácido DesoxirriboNucleico)

Todos sabemos que los elefantes solo dan a luz elefantes, las jirafas a jirafas, los perros a perros y así para cada tipo de criatura viviente. Esto se debe a una molécula llamada á*cido desoxirribonucleico* o ADN (figura 3), que contiene las instrucciones biológicas que hacen a cada especie única. El ADN se pasa de generación en generación por medio de la reproducción(Setubal & Meidanis, 1997).

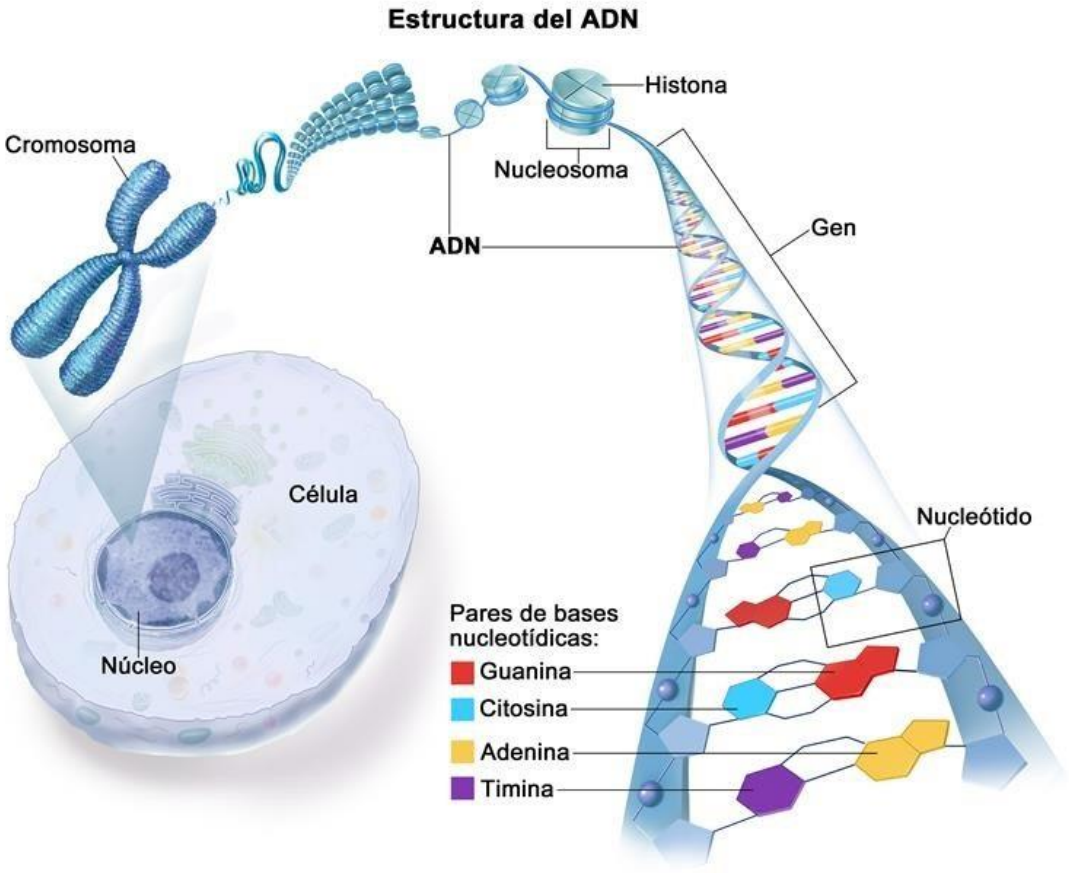

Figura 3 – Estructura del ADN.

Es un compuesto químico que contiene las instrucciones necesarias para desarrollar y dirigir las actividades de casi todos los organismos vivos. Las moléculas de ADN están compuestas por lo que se denomina doble hélice, que son dos hebras de moléculas enrolladas y apareadas.

Cromosomas

En las células eucariotas, el ADN se encuentra dentro de un área especial de la célula denominada núcleo. Debido a que la célula es muy pequeña y a que los organismos poseen muchas moléculas de ADN por célula, cada una de estas moléculas debe estar estrechamente comprimida. Esta compresión del ADN se llama *cromosoma*.

Cada hebra de ADN está formada por unidades químicas llamadas bases nucleótidas. Estas bases son Adenina (A), Timina (T), Guanina (G) y Citosina (C). El orden de estas bases determina el significado de la información codificada en esa parte de ADN, así como el orden de las letras determina el significado de una palabra. El conjunto completo de ADN nuclear de un organismo es llamado ***genoma.***

Secuenciamiento del ADN

Secuenciar simplemente significa determinar el orden exacto de las bases en una hebra de ADN. Debido a que estas bases existen en pares (solo se liga Adenina con Timina y Guanina con Citosina), al identificar una de las bases en el par, se puede determinar la otra, por esta razón nunca se reportan ambas bases del par.

Los dos protocolos clásicos de secuenciación (método *químico* y *enzimático*) comparten etapas comunes.

- Marcado: Es necesario marcar las moléculas a secuenciar radiactiva o fluorescentemente.
- Separación: Cada protocolo genera una serie de cadenas sencillas de ADN marcadas cuyos tamaños se diferencian en una única base. Estas cadenas de distintas longitudes pueden separarse por electroforesis en geles desnaturalizantes de acrilamida-bisacrilamida-urea, donde aparecen como una escalera de bandas cuya longitud varía en un único nucleótido.

Secuenciación de Maxam-Guilbert (químico)

Esta técnica (Maxam & Gilbert, 1977)consiste en romper cadenas de ADN de cadena sencilla marcadas radiactivamente con reacciones químicas específicas para cada una de las cuatro bases. Los productos de estas cuatro reacciones se resuelven, por electroforésis, en función de su tamaño en geles de poliacrilamida donde la secuencia puede leerse en base al patrón de bandas radiactivas obtenidas.

Secuenciación de Sanger (enzimático)

Este método de secuenciación(Sanger, Nicklen, & Coulson, 1977) de ADN fue diseñado por Sanger, Nicklen y Coulson también en 1977 y se conoce como método de los terminadores de cadena o dideoxi.

Para este método resulta esencial disponer de un ADN de cadena simple (molde) y un iniciador, cebadoro "primer" complementario de una región del ADN molde anterior a donde va a iniciarse la secuencia. Este cebador se utiliza como sustrato de la enzima ADN polimerasa I que va a extender la cadena copiando de forma complementaria el molde de ADN.

Secuenciación automática derivadadel método enzimático

Es una alternativa al método de Sanger. Consiste en marcar el oligo cebador o los terminadores con un compuesto fluorescente y activar la reacción de secuencia. Los productos de la reacción se detectan directamente durante la electroforésis al pasar por delante de un láser que al excitar los fluoróforos permite detectar la fluorescencia emitida.

- Secuenciación empleando cebadores fluorescentes: Se realizan cuatro reacciones de secuencia distintas en cada una de las cuales se añade el oligonucleótido o cebador marcado con una sonda fluorescente distinta y un didesoxinucleótidos o ddNTP diferente en cada una de ellas. Al finalizar las cuatro reacciones se mezclan en un único tubo.
- Secuenciación empleando terminadores fluorescentes: Se realiza una única reacción de secuencia en presencia de los cuatro ddNTPs, cada uno de ellos marcados con una sonda fluorescente distinta (figura 4).

Secuenciación de ADN empleando microarrays

Los microarrays constituyen la última línea de técnicas basadas en la interacción de cadenas complementarias de ADN. Este tipo de técnicas introducen básicamente dos nuevas innovaciones: el empleo de soportes sólidos no porosos tales como cristal que facilitan la miniaturización y la detección basada en fluorescencia y el desarrollo de métodos de síntesis in situ de oligonucleótidos a altas densidades sobre el soporte sólido.

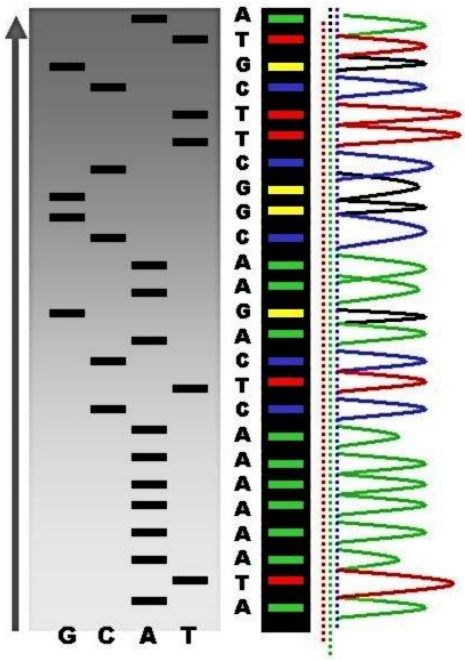

Figura 4 – Ejemplo de secuenciación de ADN automática, mediante fluorescentes.

ARN

Las moléculas de ARN o *ácido ribonucleico*, son similares a las moléculas de ADN (ver figura 5 con su comparación), con las siguientes diferencias básicas de composición y estructura (Setubal & Meidanis, 1997):

- El componente azúcar del ARN es la ribosa en vez de desoxirribosa.
- En el ARN, la Timina (T) está reemplazada por el Uracilo (U) que también liga con la Adenina.
- El ARN no forma una doble hélice.

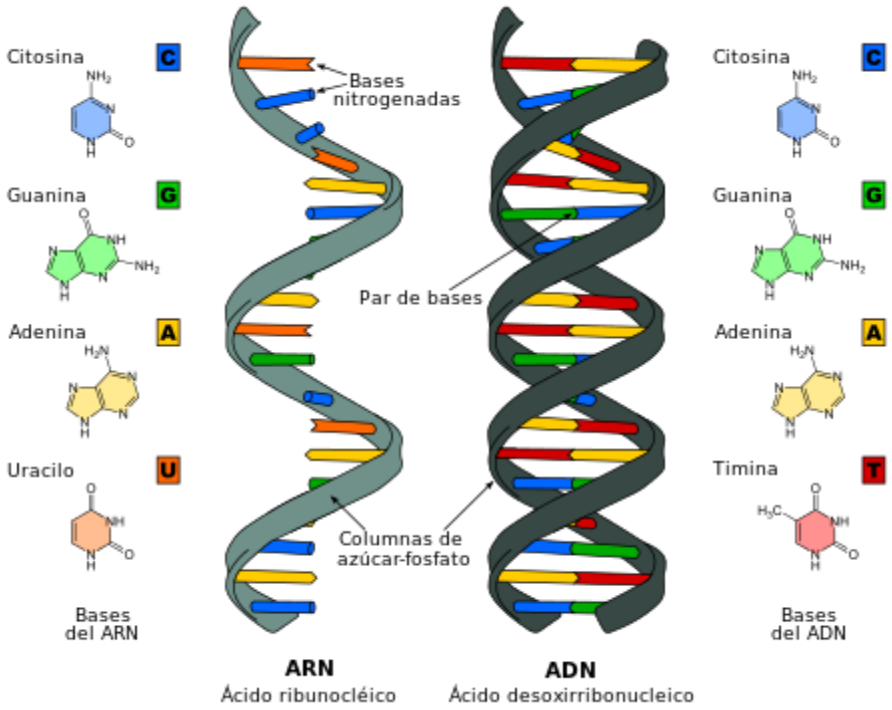

Figura 5 - ARN y ADN.

Replicación y transcripción de ADN

La replicación del ADN (figura 6) es el proceso por el cual se obtienen copias o réplicas idénticas de una molécula de ADN. La replicación es fundamental para la transferencia de la información genética de una generación a la siguiente y, por ende, es la base de la herencia. El mecanismo consiste esencialmente en la separación de las dos hebras de la doble hélice, la *ADN polimerasa* y otras enzimas sintetizan dos nuevas cadenas de ADN que son complementarias respecto a las dos cadenas originales. Durante este proceso, la ADN polimerasa reconoce una base nucleotídica no apareada de la cadena original y la combina con un nucleótido libre que tiene la base complementaria correcta. El resultado final son dos moléculas idénticas a la original.

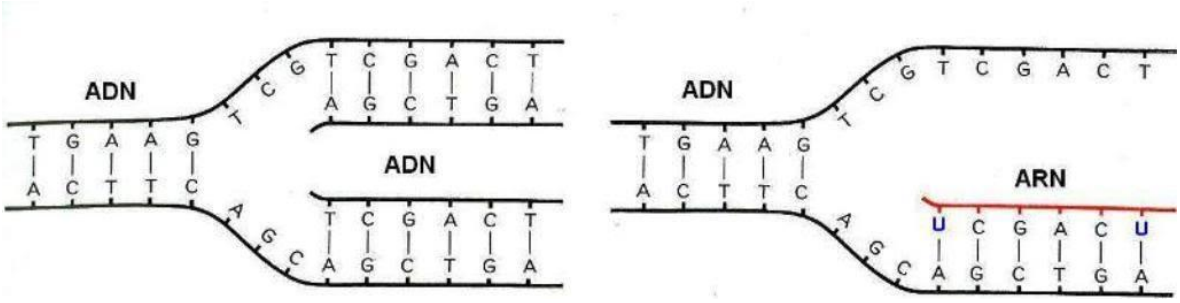

Figura 6 - Replicación (izquierda) y transcripción (derecha) del ADN.

La transcripción (figura 6) es el proceso por el cual se sintetiza un ARN usando como molde al ADN. Este mecanismo también consiste en la separación de las dos hebras de ADN, las cuales sirven de molde para la posterior síntesis de cadenas complementarias de ***ARNm*** (ARN mensajero), mediante una enzima llamada *ARN polimerasa* que sintetiza las cadenas

complementarias de ARNm Después del proceso de transcripción, el ARNm será transportado a las estructuras celulares llamadas ribosomas para guiar la síntesis de proteínas. Esta transcripción es sólo válida para células procariotas. Para eucariotas, muchos genes están compuestos por partes alternativas llamadas *intrones* y *exones*. Después de la transcripción, los intrones son quitados del ARNm. Esto significa que sólo los exones participan de la síntesis de proteínas. El *splicing alternativo* ocurre cuando el mismo ADN genómico puede dar como resultado 2 o más moléculas de ARNm diferentes dependiendo de la elección alternativa de intrones y exones, resultando generalmente en la producción de diferentes proteínas. Por esta razón, el gen completo de los cromosomas es usualmente denominado ADN genómico, y la secuencia cortada conteniendo solo exones es denominada ***ADNc*** (ADN complementario)(Setubal & Meidanis, 1997).

El ARNm viaja fuera del núcleo hacia el citoplasma de la célula donde es leído por un ribosoma, y la información es utilizada para unir las moléculas de aminoácidos en el orden correcto para formar una proteína específica. Otro tipo de ARN, llamado ***ARNt*** (ARN de transferencia) es el encargado de mapear esta información.

2.4 Ruta biológica

Una ruta biológica (o *pathway* en inglés) consiste en una serie de acciones entre moléculas de una célula que genera un cierto producto o un cambio en la misma. Tal ruta puede desencadenar la formación de nuevas moléculas, tales como proteínas. Las rutas biológicas también pueden controlar genes, o estimular a una célula a moverse (figura 7).

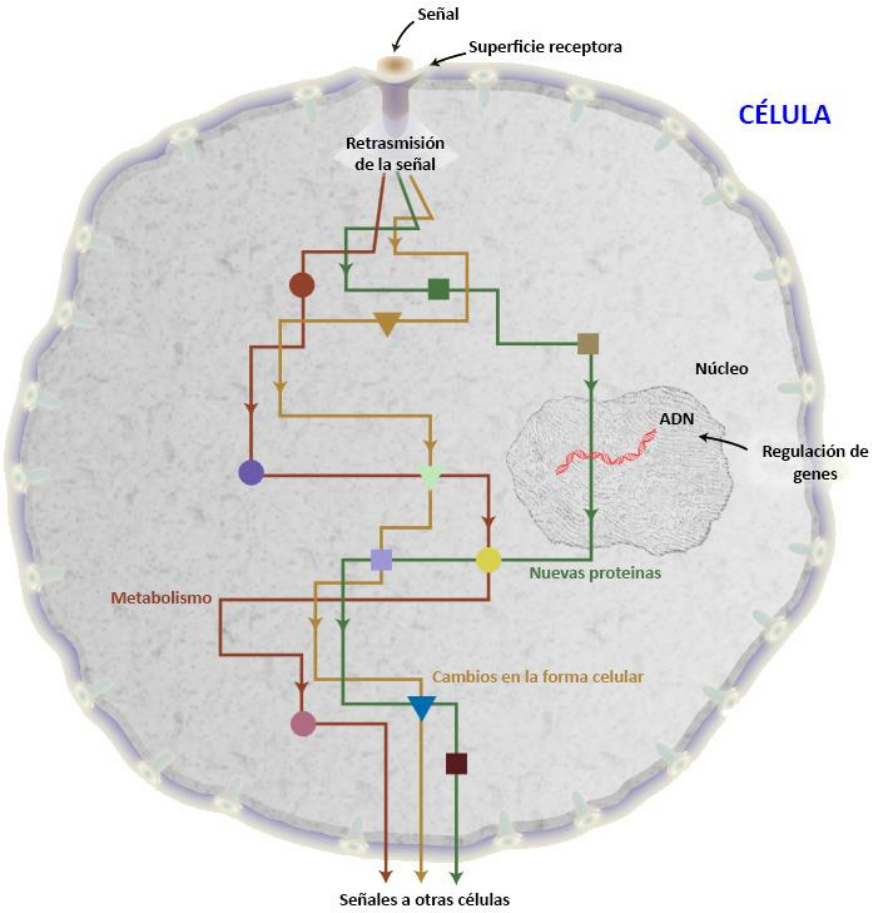

Figura 7 - Ruta biológica. Desde la recepción de una señal hasta el resultado.

Tipos de rutas biológicas

Existen varios tipos de rutas biológicas. Las más comunes están involucradas en el metabolismo, transmisión de señales y regulación de genes y se describen a continuación.

Rutas metabólicas (metabolic pathways)

Son una serie de reacciones químicas que ocurren dentro de una célula donde el químico inicial, llamado *metabolito*, es modificado por una secuencia de reacciones químicas, que son catalizadas por enzimas. Las rutas metabólicas son requeridas para la *homeostasis*, propiedad de los organismos vivos que consiste en su capacidad de mantener una condición interna estable compensando los cambios en su entorno mediante el intercambio regulado de materia y energía con el exterior. El flujo de metabolitos a lo largo de la ruta es regulado dependiendo de las necesidades de la célula. El producto final puede ser usado inmediatamente, o iniciar otra ruta metabólica o guardarse para su posterior uso. Las rutas metabólicas más significativas son aquellas que permiten el *anabolismo* y *catabolismo* de la célula, es decir, la síntesis y rompimiento de moléculas.

Rutas de transmisión de señales (signal transduction pathways o signaling pathways)

Trasladan una señal del exterior de una célula hacia su interior, o dentro de la misma célula. Las células son capaces de recibir señales específicas a través de las estructuras en su superficie, llamadas receptores. Luego de interactuar con un receptor, la señal viaja a

través dela célula donde su mensaje es transmitido por proteínas especializadas que disparan una acción específica en la célula (respuesta). Dependiendo de la célula, la respuesta puede alterar el metabolismo, forma, expresión genética o habilidad para dividirse de la misma. Por esta razón una molécula de señalización (*signaling molecule*) puede causar muchas respuestas.

Rutas de regulación de genes (genetic pathways)

Como su nombre lo indica, controlan la actividad de los genes, los "*activan*" o los "*desactivan*". Estas acciones son vitales ya que los genes producen las proteínas que, como se ha mencionado antes, son los componentes claves que llevan a cabo casi todas las tareas en nuestro cuerpo. Esas rutas son más conocidas como ***redes regulatorias de genes*** o GRN por sus siglas en inglés: *gene regulatory network*.

A cierto nivel, las células biológicas se puede considerar como "bolsas parcialmente mezcladas" de productos químicos biológicos. En el contexto de las redes regulatorias de genes, estos productos químicos son en su mayoría los ARNm y las proteínas que surgen de la expresión génica. Como hemos mencionado previamente, estos ARNm y proteínas interactúan entre sí con diferentes grados de especificidad. Algunos se difunden por toda la célula. Otros están ligados a las membranas celulares, interactuando con las moléculas en el medio. Y otros pasan a través de las membranas celulares y median en las señales de largo alcance para otras células en un organismo multicelular. Estas moléculas y sus interacciones constituyen una red regulatoria de genes. En forma general, una red regulatoria de genes es una colección de segmentos de ADN en una célula que interactúan

entre sí (indirectamente a través de su ARNm y proteínas) y con otras sustancias en la célula, con lo que regulan las tasas a las que los genes de la red se transcriben en ARNm.

Límites en el estudio de las rutas biológicas

Se han descubierto muchas rutas biológicas a partir de estudios de laboratorio sobre células cultivadas, bacterias, moscas de la fruta, ratones y otros organismos. Muchas de las rutas identificadas por estos medios son iguales o tienen una contrapartida similar en los humanos. Aun así, muchas rutas biológicas no se han encontrado. Llevaría años de investigación identificar y comprender las conexiones complejas a lo largo de todas las moléculas en todas las rutas biológicas, así como entender cómo estas rutas funcionan en conjunto.

Las rutas biológicas y las enfermedades

Se puede aprender mucho acerca de enfermedades humanas estudiando las rutas biológicas. Identificar qué genes, proteínas y otras moléculas están involucradas en una ruta biológica puede proveer pistas acerca de qué es lo que funciona mal cuando una enfermedad ataca. Por ejemplo, se puede comparar ciertas rutas biológicas en una persona sana con las mismas rutas en una persona con una enfermedad para descubrir la raíz del desorden.

Este conocimiento puede ser útil para un diagnóstico, tratamiento y prevención de una enfermedad. Cada vez más se utiliza la información acerca de rutas biológicas para desarrollar drogas nuevas y mejoradas.

Si se toma el caso del cáncer, hasta no hace mucho, los científicos esperaban que la mayoría de los tipos de cánceres fueran generados por un único error genético. De esta forma podrían ser tratados por drogas destinadas a atacar esos errores específicos. Proyectos recientes que han descifrado el genoma de células de cáncer, han detectado un conjunto de diferentes mutaciones genéticas que llevaron al mismo cáncer en distintos pacientes. Entonces, basándose en el perfil genético de un tumor particular, los pacientes podrían recibir la droga o combinación de drogas que podría funcionar para su tratamiento. La complejidad de estos descubrimientos lleva a pensar que en lugar de atacar un enemigo bien definido genéticamente, se debe afrontar el problema en una escala mayor, identificando qué rutas biológicas se ven interrumpidas por las mutaciones genéticas mencionadas.

Red de rutas biológicas

En la actualidad se conoce que las rutas biológicas son mucho más complejas de lo que se pensó en un principio. La mayoría de estas rutas no tienen un punto de comienzo ni final, de hecho, muchas de ellas no tienen límites reales y muchas veces trabajan en conjunto para realizar ciertas tareas. Cuando múltiples rutas biológicas interactúan entre sí, se forma una *red biológica*.

2.5 Sumario

En este capítulo se trataron algunos temas básicos de biología molecular. Se introdujeron los conceptos de moléculas tales como proteínas y ácidos nucleicos, y en un nivel superior se presentó primeramente el significado de las células. Se detalló el contenido genético de estas y se hizo foco en los procesos biológicos de transcripción y replicación de ADN. Como un conjunto de reacciones entre las moléculas, se definieron las rutas biológicas, y se presentaron brevemente las rutas metabólicas, de transmisión de señales y regulatorias de genes. Finalmente se hizo foco en la utilidad de conocerlas, y en las limitaciones que se presentan para poder llevar a cabo su estudio.

CAPÍTULO 3: Microarrays

Si bien todas las células del cuerpo humano contienen material genético idéntico, en cada célula no están activos los mismos genes. Estudiar qué genes están activos y cuáles inactivos en los diferentes tipos de células ayuda a entender cuando estas funcionan en condiciones normales y cómo son afectadas cuando varios genes no funcionan apropiadamente. En el pasado, estos análisis se podían hacer de a unos pocos genes a la vez. Con el desarrollo de la tecnología de microarray, sin embargo, se pueden examinar cuán activos se encuentran miles de genes en un momento dado.

Este capítulo está destinado a introducir la tecnología de microarrays, desde su funcionamiento hasta su utilidad, pasando por conceptos necesarios para entender los capítulos siguientes

3.1 Tecnología de microarrays

El concepto y metodología de los *microarrays* fue introducido por primera vez por Chang (Chang, 1983) ilustrado en una matriz de anticuerpos en 1983. La industria del "chip genético" comenzó a crecer significativamente luego de la publicación de Davis y Brown (Schena, Shalon, Davis, & Brown, 1995) en la revista Science. Con el establecimiento de compañías como Affymetrix, Agilient, Applied Microarrays, Illumina, entre otras, la

tecnología de los *microarrays* de ADN se ha convertido en una de las más sofisticadas y ampliamente usadas en la comunidad bioinformática.

Un *microarray* (también llamado chip de ADN o biochip) es una matriz de 2 dimensiones con información de grandes cantidades de material biológico. De esta forma, los datos obtenidos se usan para analizar la expresión diferencial de genes, y se pueden monitorear de manera simultánea los niveles de miles de ellos.

Más específicamente, en una matriz de vidrio, silicona, o nylon, se sitúan o "anclan" puntos (*spots*) de oligonucleótidos (fragmentos de ADN) conocidos, y en una ubicación precisa. Sobre ellos se sitúan o "hibridan" fragmentos de ADN desconocido, procedentes de tejidos o muestras de pacientes. Las bases complementarias se reconocen y se visualiza su hibridación mediante el uso de sustancias fluorescentes, aunque también se puede utilizar quimioluminiscencia o radiactividad. La miniaturización permite el anclaje, en unos pocos centímetros, de cientos de puntos de oligonucleótidos, alineados en filas y columnas, cada uno de los cuales se corresponde a una secuencia específica del ADN.

De ésta reacción se obtiene un arreglo en el que la lectura de los puntos marcados permite identificar la presencia o ausencia de los diferentes fragmentos, y componer un cuadro genómico, o huella genética, para la muestra problema. La identificación puede ser, en ocasiones, laboriosa y puede precisar de la ayuda de algoritmos y programas informáticos que ayuden a la interpretación de los resultados.

Uso de la tecnología de microarrays

La tecnología de microarray ayuda a aprender más sobre las diferentes enfermedades, incluyendo enfermedades del corazón, mentales e infecciosas. Un área intensa de la investigación de microarray en el Instituto Nacional de Salud (NIH por sus siglas en inglés *National Institute of Health*) es el estudio del cáncer. En el pasado, se han clasificado diferentes tipos de cáncer basándose en el órgano en el que los tumores se desarrollan. Con la ayuda de la tecnología de microarrays, sin embargo, se puede clasificar estos tipos de cáncer basándose en los patrones de actividad de genes en las células del tumor. De esta forma se podrían diseñar diferentes tipos de estrategias de tratamiento directamente dirigidas a cada tipo específico de cáncer. Adicionalmente, al examinar las diferencias en la actividad de genes entre células tratadas y no tratadas del tumor, se podría entender exactamente cómo diferentes terapias afectan los tumores y se podrían desarrollar tratamientos más efectivos.

Experimento de microarray

Los *microarrays* de ADN son creados por máquinas robóticas que acomodan cientos de miles de minúsculas secuencias de genes en una única diapositiva microscópica. Se posee una base de datos de más de 40000 secuencias de genes que pueden ser utilizadas para este propósito. Cuando un gen es activado, la maquinaria celular comienza a copiar ciertos segmentos de ese gen. El producto resultante es conocido como ARN mensajero (*ARNm*), que es la plantilla del cuerpo para crear proteínas. El ARNm producido por la célula es

complementario, y por lo tanto se unirá a la porción original de hebra de ADN de la cual fue copiada.

Para determinar cuáles genes se activan y cuáles no en una célula dada, se debe recolectar primero las moléculas de ARNm presentes en esa célula. Luego se etiqueta cada una de estas moléculas de ARNm utilizando una enzima transcriptasa inversa (TI o RT por sus siglas en inglés *reverse transcriptase*) que genera un ADN complementario (*ADNc*) al ARNm. Durante este proceso los nucleótidos fluorescentes son ligados al ADNc. Las muestras enfermas y normales son etiquetadas con diferentes tintes fluorescentes. A continuación, se coloca el ANDc etiquetado en una diapositiva de microarray de ADN. El ADNc etiquetado que representa el ARNm en la célula es hibridado (o ligado) a su ADN complementario sintético en la diapositiva de microarray, dejando su etiqueta fluorescente. Entonces se utiliza un escáner especial para medir la intensidad fluorescente para cada área de la diapositiva de microarray.

Si un gen particular está muy activo, produce muchas moléculas de ARNm, por lo que produce más ADNc etiquetado, que se hibrida con el ADN de la diapositiva de microarray y genera un área fluorescente muy brillante. Los genes que en cambio están muy poco activos producen menos ARNm, por lo que, menos ADNc etiquetado, que resulta en áreas fluorescentes apagadas. Si no hay fluorescencia, ninguna de las moléculas mensajeras se ha ligado al ADN, indicando que ese gen está inactivo.

Esta técnica se utiliza frecuentemente para examinar la actividad de varios genes en diferentes tiempos. A continuación se muestra un ejemplo de experimento de microarray (figura 8):

1. Las dos muestras a ser comparadas son adquiridas. Por ejemplo una muestra tratada y una no tratada con cierta droga (muestra control/afectada).

2. El ácido nucleico de interés es purificado. Por ejemplo el ARN total (nuclear y citoplasmático) es aislado por una extracción Trizol.
3. El ARN purificado es analizado por calidad y cantidad. Si el material es de una calidad aceptable y una cantidad suficiente se encuentra presente, el experimento puede proceder.
4. El producto etiquetado es generado por transcripción inversa (TI o RT por sus siglas en inglés: *Reverse Transcription*) y opcionalmente amplificado con PCR (Reacción en Cadena de Polimerasa). El ARN puede ser inversamente transcripto por *cebadores* polyT (específico para ARNm) o por *cebadores* aleatorios. El etiquetado es generalmente mediante fluorescentes. Este etiquetado puede ser directo (generalmente no utilizado) o indirecto (requiere una etapa de acoplamiento). Para *arrays* de 2 canales el acoplamiento ocurre antes de la hibridación y para *arrays* de 1 canal, luego de la hibridación.
5. Las muestras etiquetadas son mezcladas con una solución hibridada apropiada que consiste de varios compuestos químicos (SDS, SSC, sulfato dextran, etc)
6. La mezcla es desnaturalizada y agregada a los huecos del microarray. Estos huecos son sellados y el microarray hibridadose coloca en un horno o en un mezclador.
7. Luego de la hibridación, se elimina mediante un lavado (*washing*) los enlaces no específicos.
8. El microarray se seca y es escaneado por una máquina que utiliza un láser para excitar el tinte y mide los niveles de emisión con un detector.
9. La imagen se coloca en una plantilla cuadriculada y la intensidad de cada función es cuantificada.
10. Los datos crudos son normalizados.

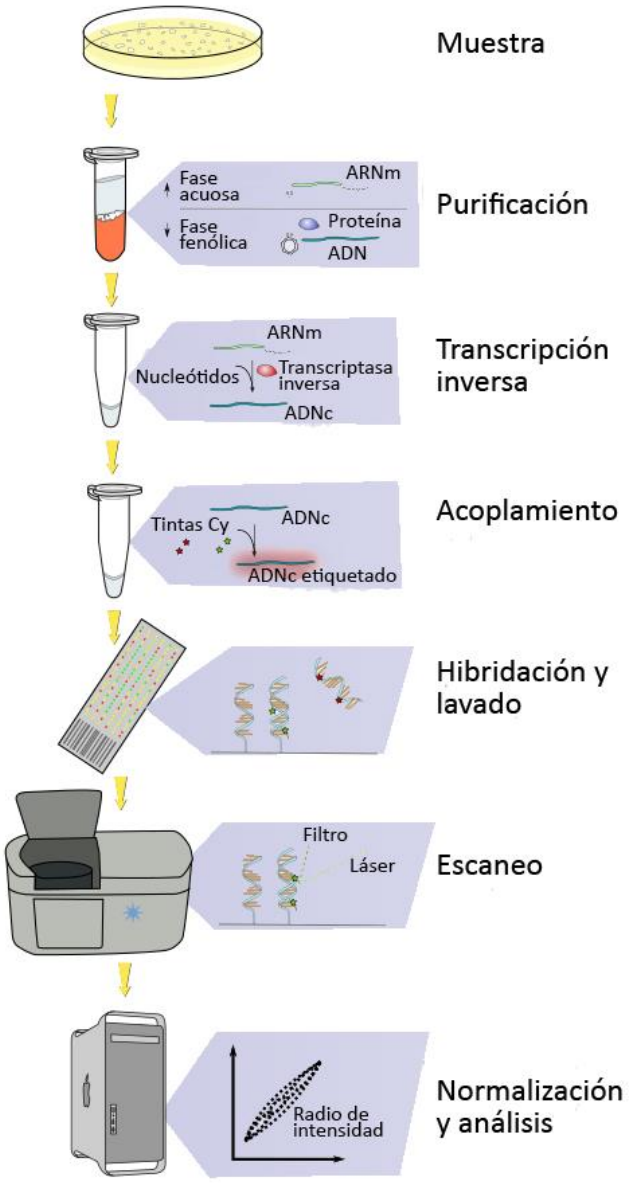

Figura 8 – Experimento de microarray de ADN.

3.2 Tipos de microarrays

Existen muchos tipos de microarrays y la distinción más amplia depende de si son espacialmente dispuestos sobre una superficie o sobre perlas/granos codificadas (*coded beads*).

En el primer caso el *array* en fase sólida tradicional es una colección de *puntos* (*spots*) ordenados microscópicos, denominados *características* (*features*), cada uno con miles de *sondas* idénticas y específicas unidas a una superficie sólida, tal como vidrio, plástico o biochip de silicona. Miles de estas características se pueden colocar en lugares conocidos en un único microarray de ADN.

En el caso del *array* de perlas o granos alternativo se trata de una colección de perlas de polietileno microscópicas, cada una con una *sonda* específica y una relación de dos o más colorantes, que no interfieren con los colorantes fluorescentes usados en la secuencia *objetivo* (*target*).

Los microarrays de ADN se pueden utilizar para detectar tanto el ADN (como en la hibridación genómica comparativa), como el ARN (más comúnmente como ADNc, o cDNA en inglés, después de la transcripción inversa) que pueden o no ser traducidos en proteínas. El proceso de medición de la expresión genética a través de ADNc se llama análisis de la expresión o de perfiles de expresión.

3.3 Fabricación de microarrays

Los microarrays se pueden fabricar de diferentes maneras, dependiendo del número de *sondas* que se examinan, costos, requisitos de personalización y el tipo de experimento a realizar. Los *arrays* o matrices pueden tener desde tan solo 10 *sondas* hasta 2,1 millones de *sondas*.

Spotted vs. in situ

Los microarrays se pueden fabricar utilizando una variedad de tecnologías, incluyendo la impresión con pins de punta fina en diapositivas de vidrio, la folitografía utilizando máscaras previamente hechas, la folitografía utilizando dispositivos dinámicos microespejos, impresión por chorro de tinta o la electroquímica en microelectrodos.

En los *microarray spotted* (manchados), las *sondas* son oligonucleótidos, ADNc o pequeños fragmentos de productos PCR que corresponden a ARNm. Las *sondas* son sintetizadas previamente a la disposición en la superficie del arreglo y luego son "*spotted*" (manchadas) en el vidrio. Un enfoque común utiliza un arreglo de cinco pins o agujas controladas por un brazo robótico que se sumerge en pozos conteniendo *sondas* de ADN y luego deposita cada *sonda* en los lugares designados de la superficie del arreglo. La "red" resultante de *sondas* representa los perfiles de ácidos nucleicos de las *sondas* preparadas y está listo para recibir los *objetivos*, el ADNc o el ARNc derivado de muestras experimentales o clínicas. Esta técnica es utilizada por los científicos investigadores alrededor del mundo para producir microarrays desde sus propios laboratorios. Estos arreglos son fácilmente personalizados

para cada experimento, debido a que los investigadores pueden elegir las *sondas* y lugares de impresión dentro del arreglo. Pueden generar sus propias muestras etiquetadas con su propio equipo. Esto provee un microarray de relativamente bajo costo que puede ser personalizado para cada estudio y evita el costo de comprar arreglos comerciales y caros que pueden representar grandes cantidades de genes en los que el investigador no está interesado. Sin embargo, las publicaciones que existen e indican este tipo de microarrays (puede ser encontrado en la literatura como *in-house spotted*) no proveen el mismo grado de sensibilidad comparado con los arreglos comerciales oligonucleótidos.

En los *microarray oligonucleótidos*, las *sondas* son secuencias cortas designadas para emparejar partes de la secuencia conocidas o predichas de *marcos de lectura abiertas* (ORF por sus siglas en inglés *open reading frames*). Estos ORF son parte de un marco de lectura que tiene el potencial para codificar una proteína o un péptido. Aunque las *sondas* oligonucleotides son generalmente utilizadas en spotted microarrays, el término "arreglo oligonucleótido" se refiere a una técnica específica de manufactura. Estos arreglos se producen por la impresión de secuencias oligonucleótidas cortas diseñadas para representar un único gen o familia de genes *splice-variants*, sintetizando esta secuencia directamente en la superficie del arreglo en lugar de depositar la secuencia intacta. Las secuencias pueden ser largas (60 *sondas* como el diseño de Agilent) o cortas (25 *sondas* producidas por Affymetrix) dependiendo del propósito deseado; mientras más largas las *sondas* más específico para genes *objetivo* específicos, mientras más cortas las *sondas*, más barato de fabricar. Una técnica utilizada para producir arreglos oligonucleótidos incluye la síntesis fotolitográfica (Affymetrix) en un sustrato sílico donde agentes enmascarantes de luz y sensibles a la luz son usados para construir una secuencia de a un nucleótido por vez a lo largo de todo el arreglo. Cada *sonda* aplicable es selectivamente "desenmascarada" previamente al *lavado* del arreglo en una solución de un único nucleótido, luego una reacción enmascarante toma lugar y el próximo conjunto de *sondas* es desenmascarado en

preparación para la exposición de un nucleótido diferente. Luego de varias repeticiones, las secuencias de cada *sonda*quedan completamente construidas.

Detección de dos canales vs. un canal

Los microarrays de *dos colores* o *dos canales* son generalmente hibridados con ADNc preparados de dos muestras que deben ser comparadas (p.e. control vs. afectados) y que son etiquetadas con dos compuestos fluorescentes distintos. Los tintes fluorescentes comúnmente utilizados para el etiquetado del ADNc incluyen Cy3 que tiene una longitud de onda de emisión de 570nm (correspondiente a la parte naranja del espectro de luz), y Cy5 con una emisión fluorescente de longitud de onda de 670nm (correspondiente a la parte roja del espectro de luz). Las dos muestras de cADN etiquetadas con Cy se mezclan y se hibridan en un único microarray que luego es escaneado para visualizar la fluorescencia de los dos tintes luego de una excitación por láser de una longitud de onda definida. Para identificar genes sobre-expresados y sub-expresados se pueden utilizar intensidades relativas a cada fluorescente.

Los microarrays oligonucleótidos generalmente llevan *sondas* de control designadas para hibridarse con *spike-ins* de ARN. Un spike-in de ARN es una transcripción de ARN utilizada para calibrar las mediciones en un experimento de microarrays de ADN, cada spike-in está diseñado para hibridar con una *sonda* de control específica sobre el arreglo *objetivo*. Para normalizar las mediciones de hibridación para las *sondas objetivo* se utiliza el grado de hibridación entre el spike-in y las *sondas de control*.Aunque niveles absolutos de expresión de genes pueden ser determinados en los arreglos de dos colores en raras ocasiones, las diferencias relativas de expresión a lo largo de diferentes puntos de una muestra y entre

muestras es el método preferible para análisis de datos del sistema de dos colores. Ejemplos de proveedores para estos microarrays son Agilent con su plataforma Dual-Mode, Eppendorf con su plataforma DualChip, y TeleChem con Arrayit.

En microarrays de *un canal* o de *un color*, los arreglos proveen datos intensos para cada sonda o conjunto de sondas, indicando el nivel relativo de hibridación con el *objetivo* etiquetado. Sin embargo, no indican verdaderamente nivel de abundancia de un gen, sino que abundancia relativa cuando se compara con otras muestras o condiciones al procesarse en el mismo experimento. Cada molécula de ARN se encuentra con el protocolo y sesgo de acuerdo al lote específico durante las fases de ampliación, etiquetado e hibridación, haciendo comparaciones entre los genes para el mismo microarray. La comparación de dos condiciones por el mismo gen requiere dos hibridaciones separadas de una tintura. Los sistemas más populares de un canal son el Gene Chip de Affymetrix, el Bead Chip de Illumina, el *One Channel Array* de Agilent, entre otros. Un punto fuerte del sistema de un color está en el hecho que una muestra anormal no puede afectar los datos crudos derivados de otras muestras, ya que cada chip es expuesto a sólo una muestra (en oposición al sistema de dos colores, en el que una muestra de baja calidad puede influir drásticamente en la precisión total de los datos). Otro beneficio es que los datos son más fácilmente comparados a arreglos de distintos experimentos siempre y cuando los efectos del lote sean tomados en cuenta.

Los microarray de un canal pueden ser la única elección en algunas situaciones. Supongamos que deben ser comparadas *n* muestras, con *n* grande, entonces el número de experimentos requeridos utilizando los microarray de dos canales se vuelven inviables, a no ser que una muestra sea usada como referencia.

3.4 Sumario

En este capítulo se trató la tecnología de microarray, desde su fabricación y su utilidad hasta una explicación para entender su funcionamiento frente a muestras genéticas, indicando los distintos tipos de tecnología y sus propósitos. De este modo se puede comprender el significado biológico de los niveles de expresión de genes obtenidos a través de estas tecnologías, lo cual es relevante para entender los datos utilizados en los siguientes capítulos de esta tesis.

CAPÍTULO 4: Biclustering y multiobjetivo

Este capítulo está destinado a introducir los conceptos básicos de *biclustering* y su variante multiobjetivo, se explican también nociones de computación evolutiva para su resolución. Se hará referencia a las matrices de expresión de genes, que son el resultado de las técnicas de *microarray* que vimos en el capítulo anterior. Como se explicó previamente, estas matrices proveen información de expresión de genes en una dimensión y condiciones o muestras en la otra, y se analizan para encontrar patrones de expresión de genes de acuerdo con las distintas condiciones en las cuales fueron medidas.

4.1 Clustering

El *clustering* es una técnica de asociación sobre un conjunto de objetos, que permite agruparlos de tal manera que los elementos del mismo grupo, denominado *cluster*, son más similares entre sí que a aquellos de otros grupos o clusters (figura 9). Es usada en minería de datos y en análisis de datos estadísticos, pero particularmente, en este trabajo de tesis, nos interesa esta técnica sobre las matrices de expresión de genes.El clustering puede ser aplicado sobre los genes o las condiciones de la misma, es decir sólo en una dirección.

La técnica de *clustering* puede ser definida como un problema de optimización y, en este sentido, el algoritmo apropiado de clustering y la elección de los parámetros dependen del conjunto de datos que se desea utilizar.

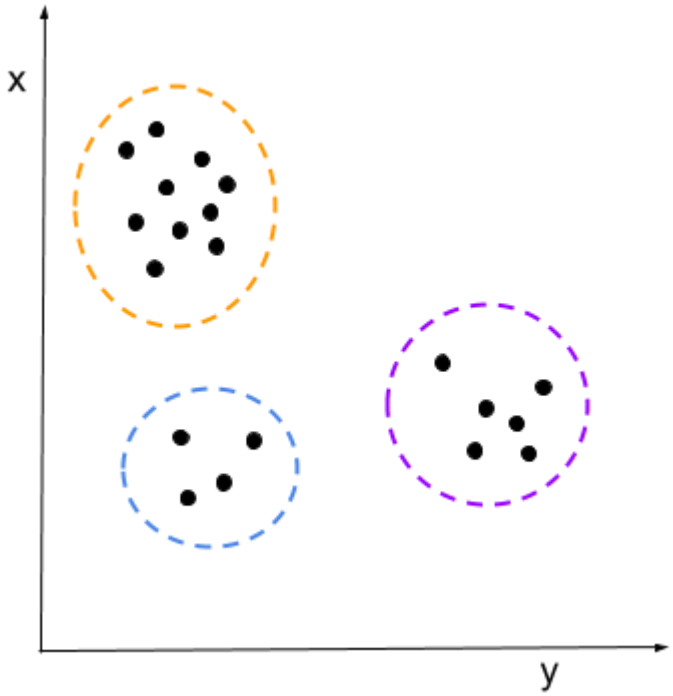

Figura 9 – Ejemplo de *clusters* en datos.

Clasificación de algoritmos de clustering

Los algoritmos de *clustering* pueden ser clasificados de la siguiente manera:

- Clustering exclusivo: los datos son agrupados de forma exclusiva, de tal forma que si un cierto dato pertenece a un cluster definido, entonces no puede ser incluido en ningún otro cluster. El algoritmo más utilizado es el de *K-Means*.

- Clustering solapado: los datos se organizan mediante una agrupación borrosa (*fuzzy*), de tal forma que cada punto o dato puede pertenecer a más de un cluster con diferentes grados de pertenencia. En este caso los datos serán asociados con un valor de pertenencia apropiado. El algoritmo más reconocido es el de *Fuzzy C-Means*.
- Clustering jerárquico: está basado en la unión entre los dos clusters más cercanos. Se comienza por la premisa de que cada dato es un cluster y luego de algunas iteraciones se llega a la cantidad final de clusters deseada. El algoritmo más utilizado lleva el mismo nombre.
- Clustering probabilístico: utiliza un enfoque probabilístico donde cada dato tiene una probabilidad determinada de pertenecer a cada cluster o clase. El algoritmo más popular de este tipo se denomina *Mixture of Gaussians*.

Dificultades de clustering

Aplicar estos algoritmos de *clustering* puede conllevar a dificultades significativas, debido a que muchos patrones de activación de genes son comunes a un grupo de estos sólo bajo ciertas condiciones experimentales, por lo que se hace casi imposible encontrar estos patrones con una simple técnica de clustering. Por esta razón se han desarrollado enfoques capaces de descubrir patrones en los datos de microarray teniendo en cuenta grupos de genes y grupos de condicionessimultáneamente: a esta técnica se la denomina *biclustering*.

4.2 Biclustering

El término biclustering fue utilizado por Cheng y Church(Cheng & Church, 2000) por primera vez en el análisis de datos de expresión de genes. Se refiere a una clase de algoritmos de clustering que realizan agrupamiento simultáneo en filas y columnas. Mientras el clustering puede ser aplicado a filas o columnas de la matriz de datos en forma separada, el biclustering realiza el clustering de estas dos dimensiones simultáneamente. El objetivo de estas técnicas es identificar grupos de genes que muestran patrones de actividad similar bajo un subconjunto específico de condiciones experimentales.

Tipos de bicluster

Existen diferentes definiciones de *bicluster*, dependiendo de su tipo (ver figura 10):

1. *Bicluster con valores constantes*: Un bicluster con valores constantes perfecto es una matriz donde todos sus valores son iguales. El algoritmo de Hartigan (Hartigan, 1972) utiliza la varianza para computar biclusters constantes, por lo que un bicluster perfecto sería aquel cuya varianza es cero. Para evitar que esto cree biclusters de una sola fila y una sola columna, Hartigan asume que hay una cantidad *K* de biclusters, cuando la matriz de datos es particionada en *K* biclusters el algoritmo termina. Cuando un algoritmo de biclustering busca un bicluster constante, la forma normal de hacerlo es reordenando las filas y columnas de la matriz para que pueda agrupar filas/columnas similares y encontrar biclusters con valores similares. Este método no soporta ruidos en los datos.
2. *Biclusters con valores constantes en filas o columnas*: Para el primer caso, con valores constantes en filas, son biclusters cuyos genes presentan un

comportamiento similar frente a las condiciones, aunque con diferentes valores de expresión para cada gen. En el caso de biclusters con valores constantes en columnas, se recogen un conjunto de condiciones, donde en cada una de ellas los genes presentan el mismo valor de expresión, pero variando de una condición a otra, aquí los genes presentan un comportamiento idéntico entre ellos. Un primer paso usual para identificar estos biclusters es normalizar las filas o las columnas de la matriz utilizando para ello la media de fila o de columna, respectivamente. Mediante esta práctica, este tipo de biclusters podrían transformarse en biclusters de valores constantes.

3. *Biclusters con valores coherentes*: Para estos casos se necesita un algoritmo más sofisticado que en los casos anteriores. Estos algoritmos pueden contener análisis de la varianza entre grupos, utilizando co-varianza entre filas y columnas. En el teorema de Cheng y Church(Cheng & Church, 2000), un bicluster es definido como un subconjunto de filas y columnas con casi el mismo puntaje, donde este puntaje de similitud es utilizado para medir la coherencia entre filas y columnas.

1.00	1.00	1.00	1.00
1.00	1.00	1.00	1.00
1.00	1.00	1.00	1.00
1.00	1.00	1.00	1.00

(A)

1.00	1.00	1.00	1.00
2.00	2.00	2.00	2.00
3.00	3.00	3.00	3.00
4.00	4.00	4.00	4.00

(B)

1.00	2.00	3.00	4.00
1.00	2.00	3.00	4.00
1.00	2.00	3.00	4.00
1.00	2.00	3.00	4.00

(C)

1.00	2.00	5.00	0.00
2.00	3.00	6.00	1.00
4.00	5.00	8.00	3.00
5.00	6.00	9.00	4.00

(D)

Figura 10 - Ejemplo de tipos de biclusters. (A) Bicluster con valores constantes. (B) Bicluster con valores constantes en filas. (C) Bicluster con valores constantes en columnas. (D) Bicluster con valores coherentes.

Definición del problema de biclustering con matrices de datos

Sea A una matriz de datos de $n$por m, es decir n filas y m columnas, donde cada elemento de A denominado a_{ij} será generalmente un valor real. En nuestro caso, por tratarse de una matriz de expresión de genes, cada elemento representa el nivel de expresión del gen i bajo la condición j. Otra forma de denotar la matriz A es con la expresión de sus filas y columnas: $I = \{i_1, i_2, \dots, i_n\}$ y $J = \{j_1, j_2, \dots, j_m\}$, siendo $A = (I, J)$. Sean $G \subseteq \mathrm{I}$ y $C \subseteq \mathrm{J}$, subconjuntos de filas y columnas de A respectivamente. Se denominará$A_{GC} = (G, C)$ a la submatriz de A con elementosa_{ij}tal que $i \in G$ y $j \in C$.

Podemos entonces definir:

- *Cluster de filas*: subconjunto de filas que exhiben comportamiento coherente o similar en todo el conjunto de columnas. Esto es, una submatriz $A_{GJ} = (G, J)$ definida sobre todo el conjunto de columnas J, donde $G = \{i_1, i_2, \dots, i_k\}$con $G \subseteq \mathrm{I}$ y $k \leq n$. Entonces A_{GJ}es una submatriz con dimensión k por m de A.
- *Cluster de columnas*: subconjunto de columnas que exhiben comportamiento coherente o similar en todo el conjunto de filas. Esto es, una submatriz$A_{IC} = (I, C)$ definida sobre todo el conjunto de filas I, donde $C = \{j_1, j_2, \dots, j_l\}\}$con $C \subseteq \mathrm{J}$ y $l \leq m$. Entonces A_{IC}es una submatriz con dimensión $n$por l de A.
- *Bicluster*: subconjunto de filas que exhiben comportamiento coherente o similar en un subconjunto de columnas. Esto es, una submatriz $A_{GC} = (G, C)$ definida en un subconjunto de filas $G = \{i_1, i_2, \dots, i_k\}$ con $G \subseteq \mathrm{I}$y $k \leq n$; y un subconjunto de columnas $C = \{j_1, j_2, \dots, j_l\}$con $C \subseteq \mathrm{J}$y $l \leq m$. Entonces A_{GC} es una submatriz con dimensión k por $l$de A. En la figura 11 se presentan ejemplos de disposición de biclusters con distintos solapamientos de genes en una matriz de datos.

El problema específico abordado por los algoritmos de biclustering puede ser definido ahora de la siguiente manera: dada una matriz de datos A, queremos identificar un conjunto de biclusters $B = (G, C)$, tal que cada bicluster B es maximal respecto de su tamaño mientras que satisface algunas características específicas de homogeneidad. Aun cuando las características exactas de homogeneidad varíen de un método a otro, es importante considerar que la varianza de cada fila en el bicluster sea relativamente alta, con el fin de capturar genes que exhiban tendencias coherentes fluctuantes bajo algún conjunto de condiciones, evitando así biclusters constantes triviales(Cheng & Church, 2000).

Supongamos que la homogeneidad $h(G, C)$está dada por la métrica residuo cuadrado promedio, mientras que el tamaño del bicluster es representado por el número de filas $f(G)$ y por el número de columnas $g(C)$; y la varianza $k(G, C)$ es la varianza de filas(Cheng & Church, 2000). Entonces una forma de definir el problema de optimización es la siguiente:

maximizar

$$f(G) = |G|$$

$$g(C) = |C|$$

$$k(G, C) = \frac{\sum_{i \in G, j \in C} (a_{ij} - a_{iC})^2}{|G||C|}$$

sujeto a

$$h(G, C) \leq \delta$$

con

$$(G, C) \in \nabla, \nabla = 2^{\{1,\dots,m\}} x 2^{\{1,\dots,n\}}$$

siendo el conjunto de todos los posibles biclusters, donde

$$h(G, C) = \frac{\sum_{i \in G, j \in C} (a_{ij} - a_{iC} - a_{Gj} + a_{GC})^2}{|G||C|}$$

es el residuo cuadrado promedio,

$$a_{iC} = \frac{\sum_{j \in C} a_{ij}}{|C|}, a_{Gj} = \frac{\sum_{i \in G} a_{ij}}{|G|}$$

son el promedio de los valores de expresión de filas y columnas de(G, C)y

$$a_{GC} = \frac{\sum_{i \in G, j \in C} a_{ij}}{|G||C|}$$

es el valor promedio de expresión sobre todas las celdas que están contenidas en el bicluster (G, C). El umbral δ definido por el usuario representa el valor máximo permitido de disimilaridad entre las celdas del bicluster. En otras palabras, el residuo cuantifica la diferencia entre el valor actual de un elemento a_{ij} y su valor esperado tal como es predicho por el correspondiente promedio de fila, promedio de columna, y promedio del bicluster. Si un bicluster tiene un residuo cuadrado promedio menor que un dado valor δ, entonces se denomina al bicluster como δ-bicluster.

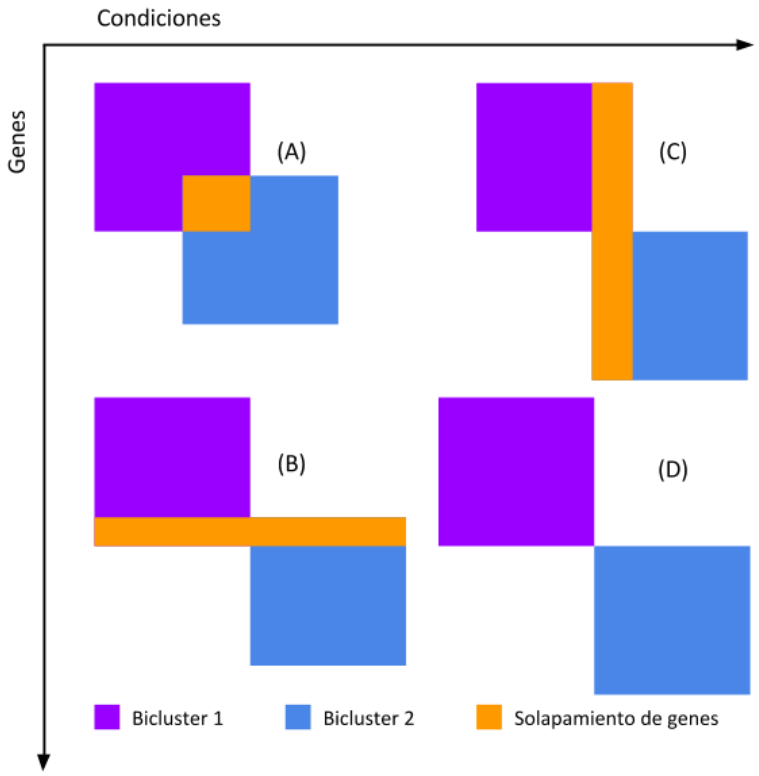

Figura 11 – Dos biclusters y distintos solapamientos de genes. (A) Comparten genes y condiciones. (B) Comparten solo genes. (C) Comparten solo condiciones. (D) No comparten genes ni condiciones.

Patrones de bicluster

La característica principal de los biclusters con valores coherentes es que los genes siguen un comportamiento similar a menudo llamado *"patrón"*. Dos modelos diferentes se pueden definir: patrones de *escalamiento* y de *corrimiento*(Aguilar-Ruiz, 2005)(Zhao, Liew, Wang, & Yan, 2012). Un bicluster B sigue un patrón de *corrimiento*cuando cada valor del gen $j$para la condición i, w_{ij}se puede obtener mediante la adición de un valor dado B_i (que permanece constante a lo largo de la i-ésima condición) y un $\Pi_j$valor típico para el gen j-

ésimo. Formalmente, un bicluster exhibe un patrón de *corrimiento* cuando sus valores se pueden describir mediante la siguiente expresión:

$$w_{ij} = \Pi_j + B_i + \xi_{ij}$$

donde w_{ij} denota el valor para el gen j bajo la condición i; B_i es el valor de desplazamiento para la i-ésima condición y ξ_{ij} representa un error. En la figura 12, se muestra un bicluster que presenta un patrón de *corrimiento*.

	c1	c2	c3	c4
g1	6	9	7	15
g2	15	18	16	24
g3	12	15	13	21

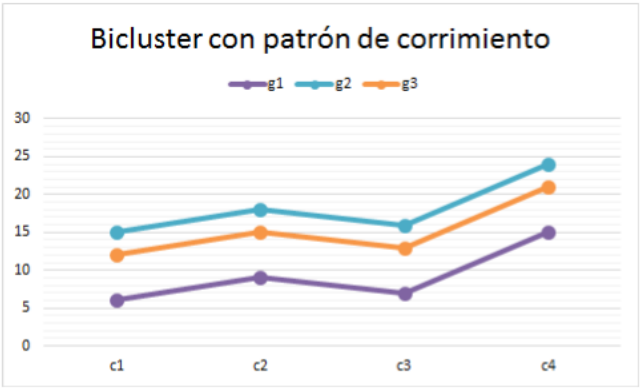

Figura 12 – Bicluster con patrón de corrimiento.

Del mismo modo, la definición de un patrón de *escalamiento* es análoga a la anterior, pero sustituyendo el valor aditivo B_i con uno multiplicativo α_i, como se muestra en la siguiente expresión:

$$w_{ij} = \Pi_j * \alpha_i + \xi_{ij}$$

donde w_{ij} denota el valor para el gen j bajo la condición i-ésima; α_i es el valor de escala para la condición i y ξ_{ij} representa un error. En la figura 13, se representa un bicluster que presenta un patrón de *escalamiento*.

	c1	c2	c3	c4
g1	8	20	12	44
g2	26	65	39	143
g3	20	50	30	110

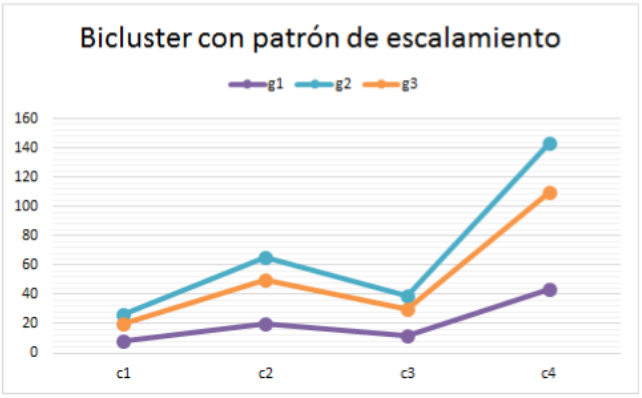

Figura 13 – Bicluster con patrón de escalamiento.

Estructuras de bicluster

Los algoritmos de biclustering pueden plantearse dos tipos distintos de objetivos: encontrar un solo *bicluster* (figura 14 A), o encontrar N biclusters, siendo N un número conocido o no de antemano. La mayoría de técnicas de biclustering asumen la existencia de varios biclusters en la matriz de datos(Hartigan, 1972)(Cheng & Church, 2000)(Getz, Levine, & Domany, 2000)(Tang, Zhang, Zhang, & Ramanathan, 2001)(Tanay, Sharan, & Shamir, 2002)(Kluger, Basri, Chang, & Gerstein, 2003)aunque otras sólo intentan encontrar uno. Estas últimas, aunque tienen la capacidad de poder localizar más de un resultado, siempre se centran en aquel que siga un determinado criterio(Ben-Dor, Chor, Karp, & Yakhini,

2002). Cuando los algoritmos asumen que existe más de un *bicluster* en la matriz de datos, como resultado se pueden obtener varios tipos de estructuras, ver figura 14:

- Biclusters con filas y columnas exclusivas (figura 14 B).
- Biclusters sin solape con estructura de tablero de ajedrez (figura 14 C).
- Biclusters con filas exclusivas (figura 14 D).
- Biclusters con columnas exclusivas (figura 14 E).
- Biclusters no solapados con estructura de árbol (figura 14 F).
- Biclusters no solapados y no exclusivos (figura 14 G).
- Biclusters solapados con estructura jerárquica (figura 14 H).
- Biclusters arbitrarios(figura 14 I).

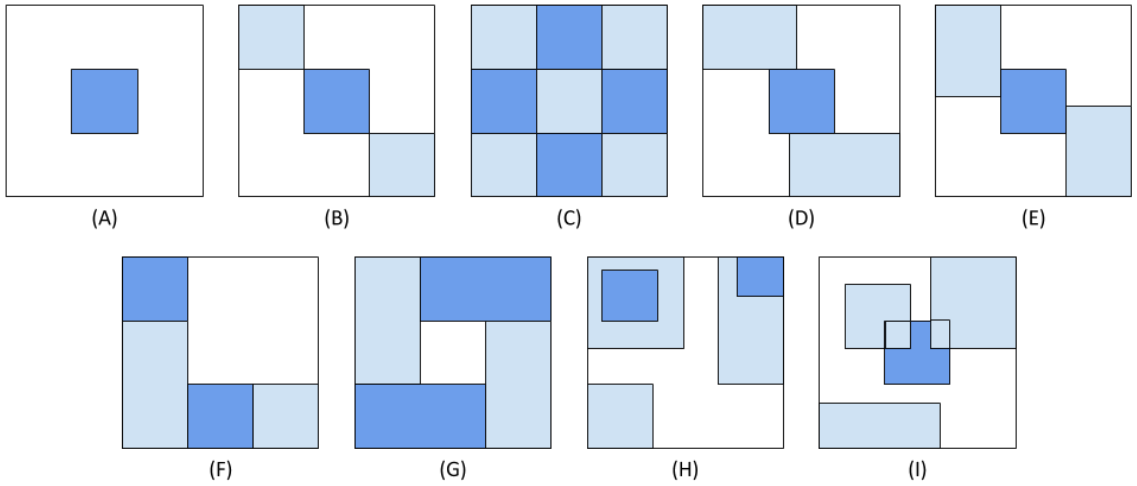

Figura 14 - Diferentes estructuras de biclusters, enumeradas previamente.

Métricas para la evaluación de los biclusters

Varianza

La varianza (VAR) determina la manera en la que los datos se extienden alrededor de un valor central, tales como la media de los valores en el bicluster B. Se puede definir como:

$$VAR(B) = \sum_{i \in I, j \in J} (b_{ij} - b_{IJ})^2$$

donde b_{ij} es el elemento en la fila i y la columna j y b_{IJ} es la media de todos los valores del bicluster. El valor de esta ecuación es 0 cuando B es un bicluster perfecto. La varianza constituye un parámetro de homogeneidad que se combina generalmente con el análisis de tamaño del bicluster en la mayoría de los métodos más simples de biclustering. En particular, la varianza por filas se utiliza generalmente como una parte de la función objetivo en muchos de ellos.

Residuo cuadrado promedio

El residuo cuadrado promedio (MSR por sus siglas en inglés *Mean Square Residue*)(Cheng & Church, 2000) cuantifica la coherencia numérica de los valores del bicluster B; cuanto menor es el valor de residuo, más fuerte es la coherencia, y mejor será la calidad de B. Se puede calcular con la siguiente fórmula:

$$MSR(B) = \frac{1}{I * J}\sum_{i=1}^{I}\sum_{j=1}^{J}(b_{ij} - b_{iJ} - b_{Ij} + b_{IJ})^2$$

donde b_{ij} es el elemento en la fila i y la columna j, b_{iJ} es la media de la columna j, b_{Ij} es la media de la fila i y b_{IJ} es la media de todos los valores en el bicluster. Un pequeño valor de MSR significa una gran coherencia entre los valores de la bicluster. Si todos los genes presentan un comportamiento idéntico en todas las condiciones, el valor de este residuo es 0.

Error Virtual

El error virtual(Pontes, Divina, Giráldez, & Aguilar-Ruiz, 2007) tiene como objetivo crear un patrón para el bicluster a fin de representar la tendencia general de todos los genes. La idea es obtener un patrón que identifica adecuadamente el comportamiento de los genes a través de todas las condiciones experimentales, independientemente de sus valores numéricos. Algunos cálculos previos se deben realizar antes de obtener el error virtual.

Dado un bicluster B con I condiciones y J genes, el *gen virtual* o patrón de comportamiento P se define como un conjunto de I elementos llamados P_i con:

$$P_i = \frac{\sum_{j \in J} b_{ij}}{J}$$

donde $b_{ij} \in \mathrm{B}, 1 \leq \mathrm{i} \leq \mathrm{I}, 1 \leq \mathrm{j} \leq \mathrm{J}$. Entonces, cada punto de P representa un valor significativo para todos los genes en una condición dada. Una vez creado el modelo, es necesario cuantificar la manera en que todos los genes se pueden ajustar a P. Con el fin de

hacer eso, tanto el patrón como el bicluster debe normalizarse, obteniendo así P' y B' de la siguiente manera:

$$b'_{ij} = \frac{b_{ij} - b_{Ij}}{b_{ij}}$$

donde $b_{ij} \in B$, $1 \leq i \leq I$, $1 \leq j \leq J$, b_{Ij} es la media de la fila i y b_{ij} es el desvío estándar de todos los valores de expresión del gen j.

$$P'_i = \frac{P_i - \overline{P}}{\sigma_P}$$

donde P_i es el valor i-ésimo del patrón (valor para la condición i), $\overline{P}$ y σ_P son la media y el desvío estándar del patrón respectivamente.

Entonces, dado un bicluster B con i condiciones y j genes, y un patrón de P con I valores, el error virtual (VE por sus siglas en inglés *Virtual Error*) del bicluster B se define como la media estandarizada de las diferencias numéricas entre cada gen y cada valor de patrón normalizado, para cada condición:

$$VE(B) = \frac{1}{I * J} \sum_{i=1}^{I} \sum_{j=1}^{J} (b'_{ij} - P'_i)$$

Un bicluster con bajo valor de VE se considera mejor que aquellos con alto valor de VE, ya que el valor de la VE disminuye cada vez que los valores de los genes son más parecidos.

Error virtual transpuesto

El error virtual transpuesto (VET por sus siglas en inglés *Virtual Error Transposed*) (Pontes, Divina, Giráldez, & Aguilar-Ruiz, 2010) constituye una mejora de su ancestro VE. Conceptualmente, esta nueva medida crea inicialmente una *condición virtual* en lugar de una estructura de *gen virtual*. Entonces, dada un bicluster B con Icondiciones y J genes, la condición virtual P se define como un conjunto de J elementos llamados P_j con:

$$P_j = \frac{\sum_{i \in I} b_{ij}}{I}$$

donde $b_{ij} \in B, 1 \leq i \leq I, 1 \leq j \leq J$. Entonces, cada elemento de P representa un valor significativo de todas las condiciones para un determinado gen. Después de este paso, las siguientes etapas son las mismas que aquellas para obtener el valor del error virtual original. B y P están estandarizados, y luego VETse calcula de manera análoga a VEpara B (sólo en este caso el vector P representa una condición virtual).

Complejidad del problema de biclustering

La complejidad de encontrar un bicluster en una matriz de datos depende de la formulación exacta del problema, y especialmente de la función utilizada para evaluar la calidad del bicluster. En particular, las variantes más interesantes de este problema, entre ellas la de detectar los biclusters con valores coherentes o encontrar el mayor δ-bicluster(Cheng & Church, 2000), son problemas NP-completos. Dichos tipos de problemas requieren un esfuerzo computacional grande o incluso obligan a la utilización de heurísticas con alta complejidad para su resolución. Esto ha motivado a los investigadores a aplicar

varias técnicas de aproximación para generar soluciones casi óptimas. Especialmente los algoritmos evolutivos son muy apropiados para abordar esta clase de problemas (Bleuler, Prelic, & Zitzler, 2004)(Divina & Aguilar-Ruiz, 2006)(Mitra & Banka, 2006).

4.3 Computación evolutiva

La computación evolutiva es un subcampo de inteligencia artificial, más específicamente de inteligencia computacional, que puede ser definida por el tipo de algoritmos que se utilizan. Estos algoritmos llamados *algoritmos evolutivos* (AE o EA por sus siglas en inglés *Evolutionary Algorithms*) están basados en la adopción de los principios Darwinianos, de aquí deriva su nombre. La computación evolutiva utiliza procesos iterativos para el crecimiento y desarrollo de una población de posibles soluciones. Esta población es luego seleccionada por medio de una búsqueda aleatoria guiada utilizando un procesamiento paralelo para obtener el fin deseado. Estos procesos están inspirados en los mecanismos biológicos de la evolución.

La evolución se produce en casi todos los organismos, como consecuencia de dos procesos primarios: la selección natural y la reproducción sexual (cruzamiento). La evolución natural es un proceso de cambio sobre una población reproductiva que contiene variedades de individuos con características heredadas y heredables. Los mismos difieren en su aptitud, la cual constituye el factor más importante al momento de obtener su éxito reproductivo. La principal idea detrás de la teoría de Darwin consiste en que las especies se crean, evolucionan y desaparecen si no se adaptan; solo los mejores, los más aptos, los que mejor se acomoden en el medio sobreviven para perpetuar sus aptitudes.

Desde el punto de vista de la computación, se puede ver un claro proceso de optimización. Se toman los individuos mejor adaptados (mejores soluciones temporales), se cruzan (mezclan), generando nuevos individuos (nuevas soluciones) que contendrán parte del código genético (información) de sus antecesores y, de esta forma, el promedio de adaptación de toda la población mejora.

Los primeros trabajos en computación evolutiva aparecieron a finales de los años 50 de la mano de Fraser (Fraser, 1958) y aprincipios de los años 60 con los trabajos de Bremermann(Bremermann, 1962) yHolland(Holland, 1962)(Holland, 1975). Sin embargo, el campo permaneció desconocido por tres décadas debido a la ausencia de una plataforma computacional poderosa y a los defectos metodológicos de los primeros métodos (Fogel, Owens, & Walsh, 1966).

Los principales beneficios de la computación evolutiva consisten en que esta técnica brinda una importante ganancia de flexibilidad y adaptabilidad a distintos problemas, en combinación con un desempeño robusto y características de búsqueda global. Constituye un concepto general que se puede adaptar para la resolución de una gran variedad de situaciones, en especial problemas de optimización difíciles de abordar con otras técnicas. Los problemas que se sugiere abordar usando los algoritmos evolutivos son aquellos entre cuyas características se encuentran alta dimensionalidad, multimodalidad, fuerte no linealidad, no diferenciabilidad, ruido y funciones dependientes del tiempo.

Los *algoritmos evolutivos* son todos los sistemas de resolución de problemas de optimización o búsqueda que emplean modelos computacionales de algún conocido mecanismo de evolución como elemento clave en su diseño e implementación. Dicho de otra forma, en general, se denomina algoritmo evolutivo a cualquier procedimiento estocástico de búsqueda basado en los principios de la evolución.

Algoritmo evolutivo

El *algoritmo evolutivo* (AE) implementa mecanismos inspirados en la evolución biológica tales como reproducción, mutación, recombinación, selección natural y supervivencia del más apto.

Las soluciones candidatas juegan el papel de individuos en una población, y la función de aptitud o *fitness function* es utilizada para la selección de los mejores individuos.

Para emular el proceso evolutivo hay dos fuerzas principales que forman parte de la transformación de los individuos: la *recombinación* y la *mutación* crean la diversidad necesaria, mientras que la *selección* actúa como el incremento de calidad.

Muchos aspectos del proceso evolutivo son estocásticos. Las partes que son cambiadas debido a la *recombinación* y *mutación* son elegidas de forma aleatoria. Por otro lado, los operadores de *selección* pueden ser determinísticos o estocásticos. En este último caso los individuos con mejor función de aptitud o *fitness function* tienen mayor probabilidad de ser seleccionados que los individuos con menor valor en dicha función, pero hasta estos últimos individuos (débiles) tienen chance de sobrevivir.

Un esquema general del algoritmo evolutivo sería el siguiente

$$t \leftarrow 0$$

$$Inicializar(P(t))$$

$$Evaluar(P(t))$$

$Hacer$

$$P'(t) \leftarrow Variar(P(t))$$

$$Evaluar(P(t))$$

$$P(t+1) \leftarrow Seleccionar(P'(t) \cup Q)$$

$$t \leftarrow t+1$$

$Mientras\ no\ se\ cumpla\ condición\ de\ término$

Donde:

- $Inicializar(P(t))$ es la creación de la población inicial $P(0)$, por lo cual usualmente se realiza asignando valores aleatorios a cada individuo.
- $P(t)$ representa la población de individuos en la generación t.
- $Evaluar(P(t))$ es la asignación de un indicador de aptitud para cada individuo de la población $P(t)$, mediante la aplicación de una función de aptitud o *fitness*.
- $P'(t)$ es la población construida a partir de la aplicación de operadores como *recombinación* y *mutación*, sobre la población $P(t)$.
- Q es un conjunto especial de individuos que pueden ser considerados para la selección. Este conjunto puede ser vacío. Además su utilización varía dependiendo del método que se utilice.
- $P(t+1)$ es la población de la generación siguiente, se obtiene a partir de la selección de la unión de la población modificada $P'(t)$ y los individuos elegibles Q, considerando la función de aptitud.

- La condición de término es un criterio que indica cuándo se debe poner fin a la búsqueda. Este criterio puede ser un nivel de convergencia, un número de generaciones máximo, o un tiempo de ejecución máximo, entre otros.

Tipos de algoritmos evolutivos

Existen tres paradigmas de algoritmos evolutivos:

1. Estrategias Evolutivas: se basan en una técnica desarrollada en sus inicios por Rechenberg (Rechenberg, 1971) y Schwefel (Schwefel, 1974), diseñada con la meta de resolver problemas de optimización discretos y continuos, principalmente experimentales y considerados difíciles. Utiliza recombinación o cruzamiento y la operación de selección, ya sea determinística o probabilística; elimina las peores soluciones de la población y no genera copia de aquellos individuos con una aptitud por debajo de la aptitud promedio.
2. Programación Evolutiva: fue introducida por Fogel (Fogel, Owens, & Walsh, 1966). Inicialmente fue diseñada como un intento de crear inteligencia artificial. El procedimiento es muy similar a las estrategias evolutivas con la diferencia de que no emplea la recombinación.
3. Algoritmos Genéticos: modelan el proceso de evolución como una sucesión de frecuentes cambios en los genes, con soluciones análogas a cromosomas. El espacio de soluciones posibles es explorado aplicando transformaciones a éstas soluciones candidatas tal y como se observa en los organismos vivientes: cruzamiento, mutación y selección. Los algoritmos genéticos constituyen el paradigma más completo de la computación evolutiva ya que en su filosofía resumen de modo natural todas las ideas fundamentales de dicho enfoque. Son muy flexibles ya que

pueden adoptar con facilidad nuevas ideas, generales o específicas, que surjan dentro del campo de la computación evolutiva. Se pueden hibridar con otros paradigmas y enfoques, aunque no tengan ninguna relación con la computación evolutiva. Por último, otra importante ventaja con respecto a las dos técnicas anteriormente presentadas consiste en que son el paradigma que cuenta con una mayor base teórica.

4.4 Algoritmos evolutivos multiobjetivo

La optimización multiobjetivo involucra a dos o más funciones objetivo que se desea optimizar simultáneamente y que se encuentran en conflicto entre sí. Existen dos objetivos (o más) en conflicto cuando no existe una única solución que simultáneamente optimice a ambos. En este caso no existe una única solución, sino que existen un número de soluciones óptimas de *Pareto*, estas soluciones se denominan así cuando ninguna de las funciones objetivo puede ser mejorada en valor, sin degradar ninguno de los valores de las otras funciones objetivo. Todas las soluciones óptimas de Pareto son consideradas igualmente buenas.

Una forma general de ver un problema multiobjetivo es la siguiente:

Maximizar o minimizar:

$$F(x) = (f_1(x), f_2(x), \ldots, f_M(x))$$

sujeto a

$$G(x) = (g_1(x), g_2(x), \ldots, g_R(x)) \geq 0$$

$$H(x) = (h_1(x), h_2(x), \ldots, h_R(x)) = 0$$

Aquí la solución corresponde un vector de variables de decisión $x = (x_1, x_2, \ldots, x_N)$ que satisfaga las restricciones impuestas por las funciones G y H, ofreciendo valores que representen un compromiso adecuado para las funciones $f_1, f_2, \ldots, f_M$.

Se define entonces un *óptimo de Pareto* a x^*, tal que para todo x en Ω(la región factible del problema), se cumple que$f_i(x) = f_i(x^*)\ \forall\ i \in \{1 \ldots M\}$, o para al menos un valor de i se cumple que $f_i(x) > f_i(x^*)$. Esto significa que no existe un vector factible que sea "mejor" que el óptimo de Pareto en alguna función objetivo sin que empeore los valores de alguna de las restantes funciones objetivo.

Asociada con la definición anterior, se introduce una relación de orden parcial denominada dominancia entre vectores solución del problema de optimización multiobjetivo. Un vector $w = (w_1, w_2, \ldots, w_N)$ domina a otro$v = (v_1, v_2, \ldots, v_N)$ si $w_i \leq v_i\ \forall\ i \in \{1 \ldots M\} \wedge\ \exists\ i\ \in \{1 \ldots M\} |\ w_i < v_i$. En este caso se nota$w \prec v$.

La resolución de un problema de optimización multiobjetivo no encuentra un único valor solución, sino que se plantea hallar un conjunto de soluciones no dominadas, de acuerdo a la definición presentada anteriormente.

El conjunto de soluciones óptimas al problema de optimización multiobjetivo se compone de los vectores factibles no dominados. Este conjunto se denomina conjunto *óptimo de Pareto* y está definido por$P^* = \{x \in \Omega \mid \nexists\ x' \in \Omega, f(x') \prec f(x)\}$. La región de puntos definida por el conjunto óptimo de Pareto en el espacio de valores de las funciones objetivo se conoce como *frente de Pareto*. Formalmente, el frente de Pareto está definido por$FP^* = \{u = (f_1(x), f_2(x), \ldots, f_M(x)) \mid x \in\ P^*\}$.

Clasificación de algoritmos evolutivos multiobjetivo

Según Coello (Coello, Van Veldhuizen, & Lamont, 2002) pueden considerarse, en general, dos tipos principales de algoritmos evolutivos multiobjetivo:

1. Los algoritmos que no incorporan el concepto de óptimo de Pareto en el mecanismo de selección del algoritmo evolutivo (p.ej., los que usan funciones de agregación lineales).
2. Los algoritmos que jerarquizan a la población de acuerdo a si un individuo es o no dominado (usando el concepto de óptimo de Pareto).

En este contexto, podemos considerar que históricamente ha habido dos generaciones de algoritmos evolutivos multiobjetivo:

1. *Primera Generación*: Caracterizada por el uso de jerarquización de Pareto y nichos. En esta generación se puede definir una nueva sub-caracterización en la que se diferencian las técnicas non-Pareto y las basadas en el óptimo Pareto. Dentro de las técnicas non-Pareto se encuentran por ejemplo las f*unciones de agregación*.
2. *Segunda Generación*: Se introduce el concepto de elitismo en dos formas principales: usando selección y una población secundaria. Los algoritmos de segunda generación enfatizan la eficiencia computacional. Se busca vencer la complejidad de la jerarquización de Pareto.

Las *funciones de agregación*, se denominan así porque integran todos los objetivos en uno solo. Se puede utilizar suma, multiplicación o cualquier otra combinación de operaciones

aritméticas. Un ejemplo de aplicación de este enfoque es mediante la suma de pesos con la forma:

$$min \sum_{i=1}^{M} p_i f_i(\overline{x})$$

donde p_i es el coeficiente de peso representando la importancia relativa de la *i*-ésima función objetivo del problema. Generalmente se asume que:

$$\sum_{i=1}^{M} p_i = 1$$

Entre las principales ventajas de esta técnica se puede mencionar que resulta muy simple implementarla y su ejecución es muy eficiente. Como desventaja podemos encontrar que, en general, las combinaciones lineales de pesos no suelen funcionar en los casos en que el frente de Pareto es cóncavo, más allá de los pesos utilizados (Das & Patvardhan, 1998). Sin embargo, los pesos pueden ser generados de tal forma que el frente de Pareto sea rotado (Jin, Olhofer, & Sendhoff, 2001) y de esta forma, los frentes de Pareto pueden ser generados de manera eficiente.

4.5 Algoritmos evolutivos multiobjetivo para biclustering

A continuación varios métodos evolutivos serán presentados cronológicamente (Carballido, Gallo, Dussaut, & Ponzoni, 2015). En todos los algoritmos, los individuos representan biclusters.

Métodos basados en MSR

En *Bleuler et al.*(Bleuler, Prelic, & Zitzler, 2004), los autores presentan el primer método que aborda biclustering de microarrays por medio de un algoritmo evolutivo. Se adopta una representación binaria de los individuos, y se utiliza mutación independiente y cruzamiento uniforme. Cada individuo representa un bicluster$B$, representado con una cadena binaria de longitud $I + J$(donde I denota el número de genes y J el número de condiciones). Un 1 en la cadena significa que el valor correspondiente se ha seleccionado para el bicluster. La función de aptitud Fse minimiza y se define en casos como sigue:

$$F(B) = \begin{cases} \dfrac{1}{|I||J|}, & MSR(B) \leq \delta \\ \dfrac{MSR(B)}{\delta}, & MSR(B) > \delta \end{cases}$$

Para la primera situación un mejor valor de aptitud, calculado utilizando sólo el tamaño del bicluster, se asigna a aquellos individuos que cumplen con la restricción de residuo. Si el bicluster tiene un residuo encima de un umbral dado, entonces se asigna un valor mayor que 1.

Diversas variantes fueron presentadas en este trabajo. Se analizan el uso de un Algoritmo Evolutivo (EA por sus siglas en inglés *Evolutionary Algorithm*) con un solo objetivo, un EA combinado con una búsqueda local (LS por sus siglas en inglés *Local Search*) (Cheng & Church, 2000) y la estrategia de LS únicamente. En el caso del EA, una novedad de la estrategia consiste en una forma de mantener la diversidad que se puede aplicar durante el procedimiento de selección. Para el caso de la EA hibridado con una estrategia de LS, consideran si el nuevo individuo producido por el procedimiento LS debe sustituir a la persona original, (enfoque de *Lamarck*) o no (enfoque *baldwiniana*). En cuanto a la LS como una estrategia independiente, proponen un método no-determinista, donde la decisión se toma en el curso de la ejecución de acuerdo con alguna probabilidad.

Otro enfoque, llamado *SEBI*por sus siglas en inglés *Secuencial Evolutive Biclustering*, fue propuesto por Divina y Aguilar-Ruiz (Divina & Aguilar-Ruiz, 2006). En este trabajo, se presenta un EA donde los individuos(biclusters) también se representan por medio de cadenas binarias. La idea principal es que el EA se ejecuta secuencialmente varias veces. De cada ejecución, el EA obtiene el mejor bicluster de acuerdo a su tamaño, la varianza por fila y factores de solapamiento. Si su MSR es inferior a un umbral dado, entonces el bicluster se añade a un archivo que ellos llaman Resultados. Siempre que este es el caso, el método sigue la pista de los elementos del bicluster con el fin de utilizar esta información para reducir al mínimo la superposición durante la siguiente ejecución de la EA. Utilizan el método de torneo para la selección y se llevan a cabo varias opciones para los operadores de recombinación. La función de aptitud combina los objetivos antes mencionados por medio de una función de agregación no-Pareto para ser minimizados como sigue:

$$F(B) = \frac{MSR(B)}{\delta} + \frac{1}{varianzaPorFilas(B)} + w_d + penalización$$

donde:

$$w_d = w_v \left(w_r \frac{\delta}{|I|} + w_c \frac{\delta}{|J|} \right)$$

teniendo w_v, w_r y $w_c$como pesos en volumen, número de filas y número de columnas en el bicluster B, respectivamente. Además

$$penalización = \sum_{i \in I, j \in J} w_p(m_{ij})$$

donde $w_p(m_{ij})$ es un peso asociado con cada elemento m_{ij} del bicluster y se define como:

$$w_p(m_{ij}) = \begin{cases} 0, & si \left| COV(m_{ij}) \right| = 0 \\ \dfrac{\sum_{k \in I, l \in J} e^{-|COV(m_{kl})|}}{e^{-|COV(m_{ij})|}}, & si \left| COV(m_{ij}) \right| > 0 \end{cases}$$

Aquí, $\left| COV(m_{ij}) \right|$ denota el número de biclusters que contienen m_{ij}. Es importante tener en cuenta que el peso $w_p(m_{ij})$ se utiliza para controlar el nivel de solapamiento entre los biclusters.

Más tarde, *Mitra y Banka*(Mitra & Banka, 2006) presentaron un Algoritmo Evolutivo Multiobjetivo (MOEA por sus siglas en inglés, *Multi-Objective Evolutionary Algorithm*) combinado con una estrategia LS (Cheng & Church, 2000). Este método constituye el primer enfoque que implementa un MOEA basado en la dominancia de Pareto para este problema. Los autores basan su trabajo en el NSGA-II (Deb, Pratap, Agarwal, & Meyarivan, 2002), y buscan biclusters con el máximo tamaño y homogeneidad. La representación del

individuo es la misma que en los métodos previamente introducidos; utiliza cruce de un solo punto uniforme, mutación de un solo bit y selección por torneo. La estrategia de LS se aplica a todos los individuos con un enfoque *lamarquiano* sugerido al comienzo de cada lazo generacional.

El *BiHEA*(Gallo, Carballido, & Ponzoni, 2009) es un EA combinado con una técnica de LS basado en el procedimiento de Cheng y Church (Cheng & Church, 2000), para orientar así la exploración y acelerar la convergencia del algoritmo evolutivo mediante el refinamiento de los cromosomas. La novedad del método de Gallo es que dos mecanismos adicionales se incorporaron en el proceso evolutivo con el fin de evitar la pérdida de buenas soluciones: un procedimiento de elitismo que mantiene los mejores biclusters, así como la diversidad en el espacio genotípico a través de las generaciones, y un proceso de recuperación que extrae las mejores soluciones de cada generación y las copia en un archivo. Este archivo es en realidad el conjunto de biclusters devueltos por el algoritmo. Aunque estos dos mecanismos parecen ser similares entre sí, hay varias diferencias entre ellos. El procedimiento elitista selecciona los mejores biclusters que no se superponen en un cierto umbral, para pasar a la siguiente generación. Estas soluciones pueden ser parte del proceso de selección de las generaciones posteriores permitiendo así la producción de nuevas soluciones basadas en estos por medio del operador de recombinación. Sin embargo, debido a las imperfecciones en el proceso de selección y de la función de aptitud, algunas buenas soluciones pueden estar siendo descartadas a través de las generaciones. Para hacer frente a este problema, se incorpora el archivo, que mantiene los mejores biclusters generados a través de todo el proceso evolutivo. Es importante remarcar que esta "*meta*" población no es parte del proceso de selección, es decir, la evolución de la población después de cada generación se controla por el proceso de recuperación sin interferir en el proceso evolutivo. En cuanto a la función de aptitud, optimiza los siguientes objetivos:

Maximizar

$$g(G, C) = |G||C|$$

$$k(G, C) = \frac{\sum_{g \in G, c \in C}(e_{gc} - e_{gC})^2}{|G||C|}$$

sujeto a

$$h(G, C) \leq \delta$$

con $(G, C) \in X,\ X = 2^{\{1 \ldots m\}} x 2^{\{1 \ldots n\}}$ siendo el conjunto de todos los biclusters, donde

$$h(G, C) = \frac{1}{|G||C|} \sum_{g \in G, c \in C} (e_{gc} - e_{gC} - e_{Gc} + e_{GC})^2$$

es el residuo cuadrado promedio (MSR), y

$$e_{gC} = \frac{1}{|C|} \sum_{c \in C} e_{gc}\ ,\ e_{Gc} = \frac{1}{|G|} \sum_{g \in G} e_{gc}$$

son la media por columna y los valores de expresión medios de (G, C), y

$$e_{GC} = \frac{1}{|G||C|} \sum_{g \in G, c \in C} e_{gc}$$

es el valor promedio de expresión sobre todas las celdas que están contenidas en el bicluster(G, C). El umbral definido por el usuario $\delta > 0$ representa la disimilitud máxima permitida dentro de las células de un bicluster. En otras palabras, el residuo cuantifica la

diferencia entre el valor real de un elemento y su valor esperado como se predijo para el correspondiente promedio por fila, promedio por columna, y promedio del bicluster.

Otro enfoque evolutivo, denominado *GABI*(por sus siglas en inglés, *Genetic Algorithm Biclustering*), se introdujo en Mukhopadhyay et al. (Mukhopadhyay, Maulik, & Bandyopadhyay, 2008). La principal diferencia con el resto de los algoritmos está en la representación de los biclusters. Aquí, cada cadena tiene dos partes, una para la agrupación de los genes y otra para la agrupación de las condiciones. Al igual que en los otros métodos, la función de aptitud utiliza MSR. En este caso, el cálculo se realiza de la siguiente manera:

$$F(B) = \frac{MSR(B)}{\delta * (1 + VAR(B))}$$

Recientemente, en Joung et al. (Joung, Kim, Shin, & Zhang, 2012) presentó un nuevo algoritmo evolutivo probabilístico, llamado *PCOBA*(por sus siglas en inglés *Probabilistic Coevolutionary Biclustering Algorithm*). La novedad de este método consiste en el uso de la información estadística global de las dos poblaciones de cooperación, de modo que la capacidad de buscar biclusters es más eficaz. La idea principal es que la estrategia co-evoluciona las dos poblaciones de biclusters para un conjunto de genes y un conjunto condiciones, y se adaptan cooperativamente entre ellos. La función de aptitud tiene por objeto reducir al mínimo el MSR, mientras que busca maximizar la varianza y el volumen del bicluster. Además, la aptitud de un individuo está determinada por el grado de cooperación entre éste y los individuos de la otra población.

Métodos basados en VE

El método propuesto en *Pontes et al.*(Pontes, Divina, Giráldez, & Aguilar-Ruiz, 2007)es un método novedoso que mejora el rendimiento de SEBI mediante la variación de la función de aptitud. La estrategia implementa la función objetivo usando la métrica de error virtual como sigue:

$$F(B) = VE(B) + w_d + penalización$$

En este caso, w_d y la penalización son los definidos previamente para el algoritmo de SEBI. Esta función de aptitud también tiene que ser minimizada. El resto del algoritmo se implementa de manera similar.

Métodos basados en VET

En Pontes et al. (Pontes, Divina, Giráldez, & Aguilar-Ruiz, 2013) se presenta el *Evo-Bexpa* (*Evolutionary Biclustering based in Expression Patterns*) que constituye el primer método biclustering en el que varios atributos de los biclusters pueden ser particularizado en términos de diferentes objetivos, de esta forma se pueden encontrar biclusters que presentan patrones de escalamiento y corrimiento (Aguilar-Ruiz, 2005)(Zhao, Liew, Wang, & Yan, 2012) simultáneamente.

Cuatro objetivos diferentes fueron individualizados en este enfoque, la medida en que un bicluster sigue un patrón de correlación perfecto, su tamaño, el nivel de solapamiento entre las diferentes soluciones y la varianza promedio. Los objetivos son considerados por

medio de la construcción de una función objetivo de agregación. Entonces, es posible especificar la influencia relativa de cada uno durante el proceso de evaluación, permitiendo así que el algoritmo sea configurable. En cuanto al primer objetivo, el VET se calcula como se explicó anteriormente. En las siguientes líneas, se describirán los otros tres términos.

En cuanto al volumen de bicluster, se define como sigue:

$$Vol(B) = \left(\frac{-ln|I|}{ln|I| + w_g}\right) + \left(\frac{-ln|J|}{ln|J| + w_c}\right)$$

donde $|I|$, $|J|$, w_g y w_c son el número de genes, el número de condiciones y los parámetros configurables para ambas dimensiones, respectivamente. La idea principal de esta ecuación es que utiliza escalas logarítmicas para que pequeños cambios en el número de filas o columnas no tengan un efecto significativo, y además separa los términos para el número de genes y el número de condiciones con el fin de evitar biclusters demasiado desequilibrados y para ser capaz de configurar cada tamaño de cada dimensión independientemente.

La superposición se controla con el siguiente término:

$$Superposición(B) = \frac{\sum_{i\in I, j\in J} W(b_{ij})}{|I||J|(n_b - 1)}$$

aquí, W es una matriz de pesos (similar a la presentada por Gasch y Eisen(Gasch & Eisen, 2002)), cuyo tamaño es el mismo que el tamaño del microarray, inicializado con valores cero al principio del algoritmo. Cada vez que un bicluster se encuentra, W se actualiza incrementando en 1 aquellos elementos que están contenidos en el bicluster. I y J se refieren a los conjuntos de filas y columnas en B, respectivamente, y $W(b_{ij})$ corresponde

al peso de b_{ij} en W. Además, n_b es el orden del bicluster solución. A grandes rasgos, este término calcula cuántas veces han aparecido los elementos de$B$ en cualquier bicluster previo, y divide este valor por el tamaño de B y el orden de la solución.

En cuanto a la varianza gen, en general se utiliza para evitar biclusters triviales, prefiriendo aquellas soluciones en las que los genes exhiben tendencias altamente fluctuantes. De acuerdo con esta idea, el término correspondiente está diseñado como sigue:

$$VarGen(B) = \frac{1}{|I||J|}\sum_{i=1}^{I}\sum_{j=1}^{J}(b_{ij} - \mu_{g_i})^2$$

como se puede observar, la varianza de un bicluster está dada por la media de las varianzas de todos los genes en el mismo.

En este enfoque se utiliza una estrategia de cobertura secuencial, donde se obtiene un único bicluster cada vez que se ejecuta el algoritmo. Entonces, si se desean *n* biclusters, el algoritmo evolutivo se debe ejecutar *n* veces.

4.6 Sumario

Este capítulo fue dedicado al concepto de biclustering, explicando en una primera fase el concepto de clustering para luego entender el de biclustering. Luego se desarrolló el concepto de computación evolutiva para sobrellevar la complejidad del problema de encontrar los biclusters. Además se explicó el concepto de un algoritmo multiobjetivo y su utilidad frente a los mono-objetivos. Finalmente se realizó un resumen de varios algoritmos evolutivos diseñados para biclustering agrupados de acuerdo a la forma de evaluar la calidad de los biclusters.

CAPÍTULO 5: Minería de datos

Este capítulo está destina a tratar los conceptos de minería de datos y texto necesarios para entender los capítulos subsiguientes. Una revisión más profunda de esta disciplina puede ser consultada en "*Modern information retrieval*" (Baeza-Yates & Ribeiro-Neto, 1999).

5.1 Minería de datos

La *minería de datos* o exploración de datos es un campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza métodos de la inteligencia artificial, aprendizaje automatizado, estadística y sistemas de bases de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. Además de la etapa de análisis en bruto, que involucra aspectos de bases de datos y de gestión de los mismos, involucra también consideraciones de inferencia, métricas de interés, consideraciones post-procesamiento de las estructuras descubiertas, visualización y hasta puede incluir actualización en línea.

La tarea de minería de datos real es el análisis automático o semi-automático de grandes volúmenes de datos para extraer patrones interesantes hasta ahora desconocidos, como

los grupos de registros de datos (*análisis de clúster*), registros poco usuales (*detección de anomalías*) y dependencias (*minería por reglas de asociación*). Esto generalmente implica el uso de técnicas de bases de datos como los índices espaciales. Ni la recolección de datos, ni la interpretación de los resultados y la información son parte de la etapa de minería de datos.

Proceso de minería de datos

Un proceso de minería de datos general sigue los siguientes pasos:

1. *Selección del conjunto de datos*: selección de variables objetivo (aquellas que se quieren predecir o calcular) y las variables independientes (que sirven para realizar algún cálculo o proceso). La selección de datos no es lo mismo que la recolección, ya que esta última no forma parte del proceso de minería de datos. Esta etapa también se la suele llamar *determinación de objetivos*.
2. *Análisis de las propiedades de los datos*: por ejemplo histogramas, diagramas de dispersión, encontrar valores atípicos o ausencia de datos (valores nulos).
3. *Transformación del conjunto de datos de entrada*: también se conoce como *preprocesamiento* de los datos, se realiza con el objetivo de preparar los datos para poder aplicar la técnica de minería de datos.
4. *Seleccionar y aplicar la técnica de minería de datos*: se construye el modelo predictivo, de clasificación o segmentación.
5. *Extracción de conocimiento*: representado por patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre las mismas.

6. *Validación y evaluación de los datos*: se debe validar el modelo, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. La interpretación de los datos es realizada por personas o especialistas en la materia.

Técnicas de minería de datos

Las técnicas de minería de datos provienen de la inteligencia artificial y la estadística, es decir son algoritmos con distinto grado de sofisticación que se aplican sobre un conjunto de datos para obtener resultados interpretables. Las técnicas más representativas son:

- *Redes Neuronales*: son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso. Se trata de un sistema de interconexión de *neuronas* en una *red* que colabora para producir un estímulo de salida. En la literatura esta técnica puede ser encontrada como *redes neuronales artificiales*, RNA o ANN por sus siglas en inglés *artificial neural network*. Una red neuronal se compone de unidades llamadas neuronas, cada una de estas neuronas recibe una serie de entradas a través de interconexiones y emite una salida. Esta salida está dada por tres funciones:
    - Una *función propagación* (o *función excitación*), que consiste en la sumatoria de todas las entradas multiplicadas por el peso de su interconexión. Este peso puede ser positivo o negativo, en el primer caso la conexión se denomina *exitatoria*, en caso contrario, *inhibitoria*.
    - Una *función de activación*, que modifica a la anterior, esta función es opcional, ya que la salida puede ser la misma función de propagación.
    - Una *función de transferencia*, que se aplica al valor devuelto por la función de activación, se utiliza para acotar el valor devuelto por dicha función. Las

más utilizadas son *función sigmoidea* (valores entre 0 y 1) y la *tangente hiperbólica* (valores entre -1 y 1).

Esta técnica parte de un conjunto de datos de entrada suficientemente significativo y su objetivo es conseguir que la red *aprenda* automáticamente las propiedades deseadas. En este sentido, el diseño de la red tiene menos que ver con cuestiones como los flujos de datos y la detección de condiciones, y más que ver con cuestiones tales como la selección del modelo de red, la de las variables a incorporar y el pre-procesamiento de la información que formará el *conjunto de entrenamiento*. El proceso por el que los parámetros de la red se adecuan a la resolución de cada problema se suele denominar *entrenamiento neuronal*.

- *Regresión Lineal*: Es la más utilizada para formar relaciones entre datos. Es rápida y eficaz pero insuficiente en aspectos multidimensionales donde pueden relacionarse más de dos variables. En estadística la regresión lineal es un método matemático que modela la relación entre una variable dependiente (Y), las variables independientes (X_i) y un término aleatorio (ε). Este modelo puede ser expresado como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon\gamma$$

Donde Y es la variable dependiente o *regresando*, las X_i son las variables independientes o *regresores*, son los parámetros que miden la influencia de cada variable.

Para poder crear un modelo de regresión lineal es necesario que se cumpla con los siguientes supuestos:

1. Que la relación entre las variables sea lineal.
2. Que los errores en la medición de las variables explicativas sean independientes entre sí.
3. Que los errores tengan varianza constante.
4. Que los errores tengan una esperanza matemática igual a cero (los errores de una misma magnitud y distinto signo son equiprobables).
5. Que el error total sea la suma de todos los errores.

Existen diferentes tipos de regresión lineal que se clasifican de acuerdo a sus parámetros:

- *Regresión lineal simple*: sólo maneja una variable independiente, por lo que sólo cuenta con dos parámetros.
- *Regresión lineal múltiple*: permite analizar la relación entre dos o más variables a través de ecuaciones, maneja varias variables independientes y cuenta con varios parámetros, este tipo de regresión es mucho más compleja que la anterior.

- *Árboles de Decisión*: es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva. Un árbol de decisión lleva a cabo un análisis a medida que éste se recorre hasta las hojas para alcanzar una decisión. Estos árboles suelen contener distintos tipos de nodos:

- *Nodos Internos*: contiene una prueba o análisis sobre algún valor de una de las propiedades.
- *Nodos de Probabilidad*: indica que debe ocurrir un evento aleatorio de acuerdo a la naturaleza del problema.
- *Nodo Hoja*: representa el valor que devolverá el árbol de decisión.
- *Ramas o Arcos*: brindan los posibles caminos que se tienen de acuerdo a la decisión tomada.

- *Modelos Estadísticos*: es una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales y en la regresión para indicar los diferentes factores que modifican la variable de respuesta. El modelo más simple es:

$$Y_{(ij)} = \mu + t_i + \xi j(i)$$

Donde Y es la variable de respuesta de interés, μ es el promedio general de la población sobre la cual se está trabajando, t es la variación que se atribuye a los niveles del factor que se está evaluando (efecto de los tratamientos). ξ es la variación de los factores no controlados (el error experimental), i es el *i*-ésimo tratamiento, j es la *j*-ésima repetición de cada tratamiento, y $j(i)$ es la variación de las unidades experimentales anidado en los tratamientos.

Los modelos estadísticos pueden ser lineales o no lineales.

- *Agrupamiento o Clustering*: es un procedimiento de agrupación de una serie de vectores a partir de ciertos criterios habitualmente de distancia, se trata de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes. El concepto de clustering es visto con más detalle en otro capítulo de esta tesis.

- *Reglas de Asociación*: se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos. Según la definición original(Agrawal, Imielinski, & Swami, 1993) el problema de minería de reglas de asociación se define como sigue: sea$I = \{i_1, i_2, \dots, i_n\}$un conjunto de n atributos binarios denominados *ítems* y sea $D = \{t_1, t_2, \dots, t_m\}$un conjunto de transacciones almacenadas en una base de datos, donde cada una de estas tiene un identificador único y contiene un subconjunto de ítems de I. Una *regla* se define como una implicación de la forma $X \rightarrow Y$, donde $X, Y \in I$, los conjuntos de ítems X e Y se denominan *antecedente* y *consecuente* respectivamente. Para seleccionar reglas interesantes del conjunto total de las reglas posibles se les agregan restricciones, las más conocidas son *soporte* y *confianza*.

  El soporte representa la proporción de transacciones en la base de datos que contiene dicho conjunto de ítems:

$$sop(X) = \frac{|X|}{|D|}$$

  La confianza representa el grado con la que una regla dada es cierta, dentro del conjunto de ítems y transacciones, puede interpretarse como un estimador $P(X|Y)$, la probabilidad de encontrar la parte derecha de una regla condicionada a que se encuentre también la parte izquierda.

$$conf(X \Rightarrow Y) = \frac{sop(X \cup Y)}{sop(X)} = \frac{|X \cup Y|}{|X|}$$

  Las reglas de asociación deben satisfacer las especificaciones del usuario en cuanto a umbrales de soporte y confianza.

Para conseguir esto el proceso de generación de reglas de asociación se realiza en dos pasos:

1. Se aplica el soporte mínimo para encontrar los conjuntos de ítems más frecuentes en la base de datos.
2. Se forman las reglas partiendo de estos conjuntos frecuentes de ítems y de la restricción de confianza mínima.

Encontrar todos los subconjuntos frecuentes de la base de datos es difícil ya que esto implica considerar todos los posibles subconjuntos de ítems (combinaciones: 2^n-1). Es posible hacer una búsqueda eficiente utilizando la propiedad *anti-monótona* (o *downward-closure*,(Zaki, 2000)), que garantiza que para un conjunto de ítems frecuente, todos sus subconjuntos también son frecuentes y, del mismo modo, para un conjunto de ítems no frecuente, todos sus subconjuntos deben ser no frecuentes. Explotando esta propiedad se han diseñado algoritmos eficientes (*Apriori* y *Eclat*) para encontrar los ítems frecuentes.

Tipos de minería de datos

Según el objetivo del análisis de los datos, los algoritmos utilizados se clasifican en supervisados y no supervisados(Weiss & Indurkhya, 1998).

- *Algoritmos Supervisados*: también llamados predictivos, predicen un dato o conjunto de datos, desconocido a priori, a partir de otros conocidos. Estos algoritmos requieren realimentación explícita, lo cual usualmente se obtiene de los usuarios, quienes indican la relevancia de cada documento.

- *Algoritmos no Supervisados*: o de descubrimiento del conocimiento, descubren patrones y tendencias en los datos. Aplica un tipo de relevancia conocida como *relevancia ciega* que presume relevantes a los primeros *n* documentos que se obtienen.
- *Algoritmos Semi-Supervisados*: estos algoritmos infieren la relevancia de los documentos, por ejemplo monitoreando a los usuarios y sus comportamientos sobre los documentos.

5.2 Minería de texto

A grandes rasgos, la *minería de texto* es el análisis de los datos contenidos en texto de lenguaje natural, trabaja por medio de la trasposición de palabras y frases en datos sin estructura hacia valores numéricos que pueden ser enlazados con datos estructurados en una base de datos y analizados con las técnicas tradicionales de minería de datos.

La *minería de textos*, también es referida como *minería de datos de texto*, o *análisis de textos*, y es el proceso de derivar información de alta calidad desde uno o varios textos. La información de alta calidad es típicamente encontrada a partir de patrones y tendencias por medio de medias como ocurre en el *aprendizaje de patrones estadísticos*. La minería de texto generalmente involucra el proceso de estructurar el texto de entrada, ya sea por medio de *parsing* o utilización de características lingüísticas, para luego encontrar patrones en esta estructuración del texto y finalmente proporcionar o facilitar la interpretación de éstos. Las tareas típicas de minería de texto son la *categorización de texto*, *clustering de texto*, *extracción de concepto o entidad*, *producción de taxonomías*, *análisis de sentimiento*, *sumarización de documentos* y *modelado de relación de entidades* (es decir encontrar relaciones entre entidades conocidas).

5.3 Sumario

En este capítulo se introdujo brevemente los conceptos básicos sobre minería de datos y sus diferentes técnicas junto con sus objetivos. Por último se explicó el concepto de minería de textos dentro del campo de la minería de datos.

CAPÍTULO 6: Inferencia de redes de rutas biológicas

Este capítulo está destinado a introducir las principales técnicas para la detección de diafonía (*crosstalk*) entre rutas biológicas e inferencia de redes de rutas biológicas conocidas en la actualidad y describe las contribuciones realizadas en esta temática en el marco la presente tesis.

6.1 Redes de rutas biológicas

Para estudiar una enfermedad al nivel de sistema, los datos de microarray de ADN son comúnmente utilizados para proveer una comparación de la expresión de patrones de genes en condiciones de control vs. afectados. Debido a que esta comparación usualmente revela un gran número de genes expresados diferencialmente, es difícil, si no imposible, analizar cada gen individualmente.

En la literatura los métodos que identifican un conjunto de genes relevantes biológicamente son llamados "*de enriquecimiento de rutas biológicas*" o "*de enriquecimiento de conjuntos de genes*" ("pathway enrichment" o "gene set enrichment" en inglés). En los últimos años han surgido nuevos métodos guiados por críticas y limitaciones sobre los existentes.

6.2 Métodos relevantes para redes de rutas biológicas

Los métodos existentes en la literatura pueden tener diversos objetivos, algunos proponen inferir interconexiones o interacciones entre rutas biológicas, mientras otros aluden a encontrar rutas diferencialmente expresadas en base alos datos proporcionados. Para llegar a resultados estos métodos pueden trabajar con datos de microarray o simplemente con información topológica de las rutas biológicas otorgado por bases de datos como KEGG (Kanehisa, Goto, Sato, Kawashima, Furumichi, & Tanabe, 2014)(Kanehisa & Goto, 2000), e incluso pueden usar una combinación de ambos. Dentro de los métodos conocidos se encuentran algunos que utilizan minería de datos, mientras que otros se basan en cálculos matemáticos y estadísticos. A continuación se detallarán los métodos más relevantes que ayudaron a la creación de uno nuevo.

Rutas biológicas diferencialmente expresadas

La selección de genes expresados diferencialmente ayuda a asociar fenotipos biológicos con sus mecanismos moleculares subyacentes, por lo que proveen una visión de la función biológica.

Una de las aplicaciones más utilizada de la tecnología de microarray, es la identificación de genes expresados diferencialmente en dos condiciones. El enfoque estadístico más común es cuantificar el interés de cada gen con un *p-value*(Irizarry, Wang, Zhou, & Speed, 2009), este valor se ajusta para múltiples comparaciones, se elige un corte apropiado y se crea una lista de genes candidatos. Este enfoque es criticado por carecer o ignorar el

conocimiento biológico de cómo los genes trabajan en conjunto (rutas biológicas). Los métodos que serán resumidos a continuación se basan en estas premisas. Estos métodos son comúnmente utilizados para encontrarle un sentido biológico a la abundante cantidad de datos de genes expresados diferencialmente.

GSEA

Comenzaremos por el más mencionado en la literatura. GSEA (Subramanian, y otros, 2005) debe sus siglas en inglés a "Análisis Enriquecido de Conjunto de Genes" (o *Gene Set Enrichment Analysis*). Toma como base un experimento típico de microarray, es el caso de muestras pertenecientes a dos clases: control vs. afectados, tumores resistentes vs. sensibles a una droga, etc. Estos genes son ordenados en una lista por rangos denominada L, de acuerdo a su expresión diferencial. El verdadero desafío, como se ha dicho previamente, es darle significado biológico a esta lista.

El GSEA requiere que se defina un conjunto de genes, llamado S, *a priori*, esto es previamente a la ejecución de su algoritmo. El objetivo del GSEA es entonces, determinar si los miembros de S están distribuidos aleatoriamente en L o están en su mayoría al principio o al final de la lista. Este algoritmo sigue 3 pasos:

1. Cálculo del Puntaje de Enriquecimiento (ES por sus siglas en inglés: *enrichment score*). El ES refleja el grado con el cual el conjunto S está sobre-representado en los extremos (principio o final) de la lista L. El puntaje es calculado recorriendo la lista L, incrementado una suma estadística cuando encontramos un gen en S y decrementando cuando encontramos un gen que no está en S. Esta medición se

corresponde con una estadística ponderada tipo Kolmogorov-Smirnov (Lilliefors, 1967).

2. Estimación del nivel de significancia de ES. Se estima la significancia estadística (*p-value* nominal) de ES utilizando un procedimiento de prueba de permutación basada en fenotipo que preserva la estructura de correlación compleja de los datos de expresión genética. Específicamente, se permutan las etiquetas de fenotipo y se recalcula el ES para los datos permutados, esto es, se genera una distribución nula para el ES. Luego se calcula el p-value con esta distribución nula. Lo que se rescata es la permutación de etiquetas y no la permutación de genes, permitiendo así conservar las correlaciones gen-gen y proporcionar una evaluación biológicamente razonable.
3. Ajuste para múltiples hipótesis. Cuando una base de datos de conjuntos de genes es evaluada, se ajusta el nivel de significancia estimada. En primer lugar, se normaliza los ES para cada conjunto de genes teniendo en cuenta el tamaño del conjunto, dando lugar a un puntaje normalizado NES. Luego controlan la proporción de falsos positivos mediante el cálculo de la tasa de falso descubrimiento FDR, correspondiente a cada NES.

Este método es ampliamente utilizado como criticado. Se encuentra en la clasificación de métodos llamado Puntaje Funcional de Clases o FCS(Goeman, Van De Geer, De Kort, & Van Houwelingen, 2004) por sus siglas en inglés (*Functional Class Score*), que básicamente considera a un conjunto de genes o ruta biológica como una clase y le aplica un puntaje. Se argumenta su limitación con respecto a que cada clase o conjunto de genes es vista de manera independiente de las otras (Draghici, y otros, 2007), es decir no utiliza las relaciones de las rutas o conjuntos de genes para su interpretación biológica. Más aún toma a cada gen perteneciente a la ruta biológica con el mismo peso, sin tener en cuenta las distintas importancias de los genes dentro de la ruta. Los métodos a continuación

proponen tener en cuenta estas relaciones entre conjuntos de genes para un mayor entendimiento de los resultados

SEPEA

Este método llamado Análisis Enriquecido de Rutas Biológicas Estructuralmente Mejorado (*Structurally Enhanced Pathway Enrichment Analysis*) (Thomas, Gohlke, Stopper, Parham, & Portier, 2009) propone tener en cuenta la importancia del gen dentro de la ruta mediante la implementación de un HER (*Heavy Ends Rule*), o Regla de Peso en los Extremos, básicamente le dan mayor importancia o peso a los componentes iniciales y finales de una red biológica. Además implementan una Regla de Distancia, DR (*Distance Rule*), implicando darle importancia a los genes que están cercanamente conectados y siguen un flujo dentro de la red.

Esto hace que el método para obtener el puntaje de la ruta biológica se modifique con respecto al GSEA. Ahora posee dos componentes de acuerdo a las dos reglas mencionadas.

1. El primer componente está relacionado con la regla HER y tendrá un valor alto cuando una combinación de los genes más "importantes" estén diferencialmente representados en el experimento.
2. El segundo componente está relacionado con la regla DR y tendrá un valor alto cuando los genes diferencialmente representados en el experimento estén situados o interconectados muy juntos en la ruta biológica.

Estos dos puntajes son luego normalizados y sumados. La normalización se realiza mediante el cálculo de la media y desvío estándar. Una vez obtenido este puntaje para

cada ruta biológica, se computa un *p-value* aleatorizando los datos(Thomas, Gohlke, Stopper, Parham, & Portier, 2009).

SPIA

Las siglas provienen de Análisis de Impacto de Pathway de Señalización (*Signaling Pathway Impact Analysis*) (Draghici, y otros, 2007), y afirma que tanto los métodos FSC como los ORA, o Análisis de Sobre-Representación (*Over-Representation Analysis*) poseen grandes limitaciones que son tenidas en cuenta por este método.

Se propone un enfoque diferente mediante un Factor de Impacto (FI), que se calcula para cada conjunto de genes o ruta biológica, incorporando parámetros como topología de la ruta biológica. Este IF es calculado por la suma de 2 términos:

1. El primero es un término probabilístico que captura la importancia de la ruta biológica desde la perspectiva del conjunto de genes que contiene. Este valor corresponde a la probabilidad de obtener un valor estadístico utilizado al menos tan extremo como el observado, cuando la hipótesis nula es verdadera.
2. El segundo término es funcional, depende de la identidad de los genes específicos que se expresan diferencialmente, teniendo en cuenta las interacciones descritas en la ruta biológica (topología). En esencia es la suma de los factores de perturbación (PF) de los genes dentro de la ruta biológica. El factor de perturbación de un gen abarca tanto la información de expresión genética del experimento como la importancia del gen dentro de la ruta biológica. Para esto último se utiliza un rango similar al de google (Page, Brin, Motwani, & Winograd, 1998), es decir un gen es importante dentro de la ruta biológica si muchos genes apuntan hacia él.

Este factor de impacto es luego normalizado por la cantidad de genes diferenciales que la ruta biológica tenga. Si bien este método logra utilizar la información de la topología interna de la ruta biológica para un mayor entendimiento de los resultados de un experimento de microarray, sigue sin tener en cuenta las relaciones entre las vías biológicas.

Detección de diafonía entre rutas biológicas

Así como hay métodos que buscan darle sentido biológico a los resultados de datos de microarray utilizando rutas biológicas, hay otros métodos que buscan encontrar vínculos entre estas rutas para un mejor entendimiento del funcionamiento del sistema. Estas interacciones son comúnmente referidas en la literatura como diafonía ("*crosstalk*") entre rutas.

Function Based Analysis

Este método (Hsu & Yang , 2012) propone identificar las relaciones entre rutas biológicas por medio de las relaciones funcionales entre estas, basándose en anotaciones de Ontología de Genes o GO (Ashburner, y otros, 2000) por sus siglas en inglés. Posee un enfoque basado en funciones (FBA por sus siglas en inglés *Function Based Analysis*) para identificar similaridad funcional entre rutas utilizando Ontología de Genes de los componentes de las rutas.

Los datos de las rutas utilizados por los autores del método son bajados de la base de datos de PID (Schaefer, y otros, 2009). En GO la mayoría de los genes tienen asignados términos o

anotaciones que generalmente se basan en las rutas biológicas de los que forman parte. Por esta razón, las anotaciones de una ruta pueden ser inferidas a partir de las de sus componentes.

La FBA propuesta tiene 2 pasos:

1. Inferir los términos representativos de cada ruta: debido a que los componentes de la misma pueden tener numerosos términos de GO, todos éstos no son relevantes para describir la función de la ruta. Utilizan el test de Fischer para identificar los términos enriquecidos de un conjunto de genes dado (ruta).
2. Calcular la similaridad entre rutas: en principio, las rutas relacionadas deberían compartir anotaciones de GO, pero no tuvieron en cuenta solo eso, sino que también el contenido del término en común y la cantidad de los mismos entre dos rutas. Utilizaron un esquema de pesos de términos de GO por la frecuencia en la que estos aparecían en las anotaciones de la ruta y en el total de genes del humano (para normalizar). Computan luego una función de similitud entre las rutas, llamada *funSim*, utilizando el coseno. Mientras más alto es el resultado de *funSim*, más relacionado está el par de rutas.

La fuerza de este método está en que no utilizan solamente similitud estadística sino que agregan información biológica curada. El problema que se presenta, como en la mayoría, está en que es casi imposible encontrar relaciones entre rutas que no poseen genes en común.

Edge Ontology

Si bien no se trata de un método, cabe destacar el trabajo de los autores (Lu, y otros, 2007), quienes proponen una ontología para los vínculos entre rutas biológicas en su trabajo. Su motivación son las inconsistencias encontradas en la literatura. Como se ha mencionado, los límites de las rutas biológicas no están bien definidos, esto conlleva a contradicciones entre bases de datos comúnmente utilizadas. Proponen que se utilice una ontología para las conexiones entre las moléculas a fin de que sea más sencillo encontrar interacciones entre rutas biológicas, debido a que en las bases de datos estas rutas son estudiadas aisladamente del resto.

Detección de rutas biológicas enriquecidas con datos de expresión de genes

Existen métodos que no sólo buscan conexiones entre las rutas biológicas sino que también quieren que la información de un experimento de microarray, por ejemplo con respecto a una enfermedad, se vea volcado en estas conexiones y rutas biológicas. A continuación se detallan algunos métodos en la literatura que intentan referirse a estos problemas en conjunto.

Análisis de Significancia de los Enlaces (Significance Analysis of Links)

Los autores (Francesconi, y otros, 2008) utilizan la selección de genes significativos mediante ANOVA y la construcción de una red de rutas biológicas. Para la red utilizan un nodo por ruta, tal y como se lo menciona en la base de datos KEGG, para cada uno de estos nodos una característica es asociada, indicando si está sobre-representado (involucrado

significativamente), sub-representado (significativamente no representado) o no representado con respecto a los genes significativos. La misma clasificación se utiliza en las interacciones entre rutas, utilizando los genes en común o solapados. Esta última afirmación es la propia debilidad del método, ya que no es capaz de encontrar interacciones o relaciones entre rutas que no posean genes en común.

Método basado en Red de Proteínas

Existen métodos, como el de Liu (Liu, Wang , Zhang , & Chen , 2010) que primero construyen una red de interacción Proteína-Proteína para luego ser mapeada a una red entre rutas biológicas. Esta red de proteínas y su mapeo son basados en la información de la base de datos de KEGG. En este método, la evaluación presentada por los autores se focaliza en las proteínas y rutas que se conoce tienen relación con la enfermedad de Alzheimer.

Una vez formada la red, los enlaces son ponderados, integrando expresión y co-expresión de datos de expresión genética de la enfermedad mencionada, tomados en 6 regiones del cerebro. Con esta ponderación se evalúa si la relación entre rutas es significativa.

Además los autores realizan una matriz de similitud entre las rutas biológicas para luego determinar clusters, basándose en la significancia de la interacción. De estos clusters se identifican las rutas con relaciones más significativas. Nuevamente los enlaces son dados por el solapamiento de genes.

PathNet

Este algoritmo (Dutta, Wallqvist, & Reifman, 2012) utiliza la información de conectividad entre las rutas biológicas, hasta las no evidentes para identificar información biológica importante de los experimentos de microarray. PathNet, llamado así por utilizar la información de redes de rutas biológicas (*Pathway based Networks*), considera tanto la expresión diferencial de los genes como la de sus vecinos dentro de la ruta biológica, para fortalecer la evidencia de que dicha ruta está implicada en las condiciones biológicas que caracterizan el experimento. Además, como complemento, se utiliza la información topológica para encontrar asociaciones contextuales entre las rutas.

Este método incorpora las conexiones entre rutas biológicas utilizando la información de la base de datos KEGG, para crear lo que ellos llaman *pooled pathway*, que es una gran ruta biológica agrupada. Para el análisis enriquecido, se identifican los genes diferenciales dadas dos condiciones (p.e. control vs. afectados) en datos de expresión de genes, esta evidencia es denominada *directa* por los autores. Luego se identifica la asociación de vecinos de cada gen con respecto a los datos de expresión de genes, evidencia llamada *indirecta*. Estos vecinos son encontrados a partir del *pooled pathway*. Estas dos evidencias son combinadas, para realizar la significación estadística sobre la misma. Mediante la prueba hipergeométrica, este método descubre las rutas biológicas asociadas a la enfermedad.

El *pooled pathway* es creado a partir de los datos proporcionados por KEGG, creando así una matriz de adyacencia llamada A. Esta matriz corresponde sus dimensiones a los genes que se encuentran en todas las rutas biológicas y el contenido a si hay o no interacciones entre ellos. El elemento diagonal de la matriz fue dejado en 0 para excluir la interacción propia. Esta matriz permite no sólo ver conexiones entre rutas que comparten genes entre

sí, sino también conexiones de rutas que poseen genes no solapados y conectados entre ellos.

Si bien resulta muy interesante el concepto utilizado por los autores de PathNet sobre las conectividades entre las rutas biológicas como relevantes para el experimento, este método no utiliza toda la información que brinda KEGG.

PANA

Este método (Ponzoni, y otros, 2014) presenta una nueva metodología de cálculo para estudiar las interconexiones funcionales entre los elementos moleculares de un sistema biológico. Este enfoque utiliza mediciones genómicas de alto rendimiento y un esquema de anotación funcional para extraer un perfil de actividad de cada ruta biológica, seguido por métodos de aprendizaje automático para inferir las relaciones entre estos perfiles funcionales (figura 15). El resultado es una red global, interconectada de rutas biológicas que representa el *crosstalk* funcional del sistema molecular.

Como primer paso, PANA (*PAthway Network Analysis*), mapea los datos transcriptómicos a una base de datos de rutas biológicas para generar un conjunto de matrices de expresión (una matriz por ruta biológica). A cada una de estas matrices se le aplica PCA (*principal component analysis*) para comprimir la información a un número reducido de perfiles de expresión. Los perfiles de las rutas biológicas (PP en la literatura por sus siglas en inglés: *pathway profile*) son los puntajes de los componentes principales del PCA, seleccionados con un umbral dado.

El segundo paso consiste en obtener un conjunto de reglas de asociación que establecen conexiones de a pares entre los PP. Reglas directas y opuestas son extraídas (correlación positiva y negativa). La calidad del valor de la regla está determinada por su precisión.

Un aspecto interesante de este método es que ha encontrado enlaces entre rutas biológicas aunque no contengan genes en común, lo que presenta generalmente una limitación en los otros métodos.

Los autores comparan los resultados encontrados por este método con los encontrados por PathNet en su publicación. El método encuentra las asociaciones más relevantes reportadas por Dutta et al. y además, encuentra asociaciones nuevas, algunas de ellas sin genes en común.

PANA supera los supuestos que se han mencionado, tanto enlaces sin contener genes en común como la suposición de que los genes de una ruta contribuyen de igual manera a su actividad. Este enfoque considera la activación de una ruta biológica como un cambio coordinado y relevante de niveles de expresión de *algunos* de sus genes, se define una "firma" o "perfil" de una ruta biológica que representa los patrones de regulación transcripcionales más importantes, y éstos son utilizados para encontrar conexiones entre rutas biológicas, es decir no son necesarias proteínas o genes compartidos entre rutas. Sin embargo este método no utiliza información topológica para encontrar las diafonías.

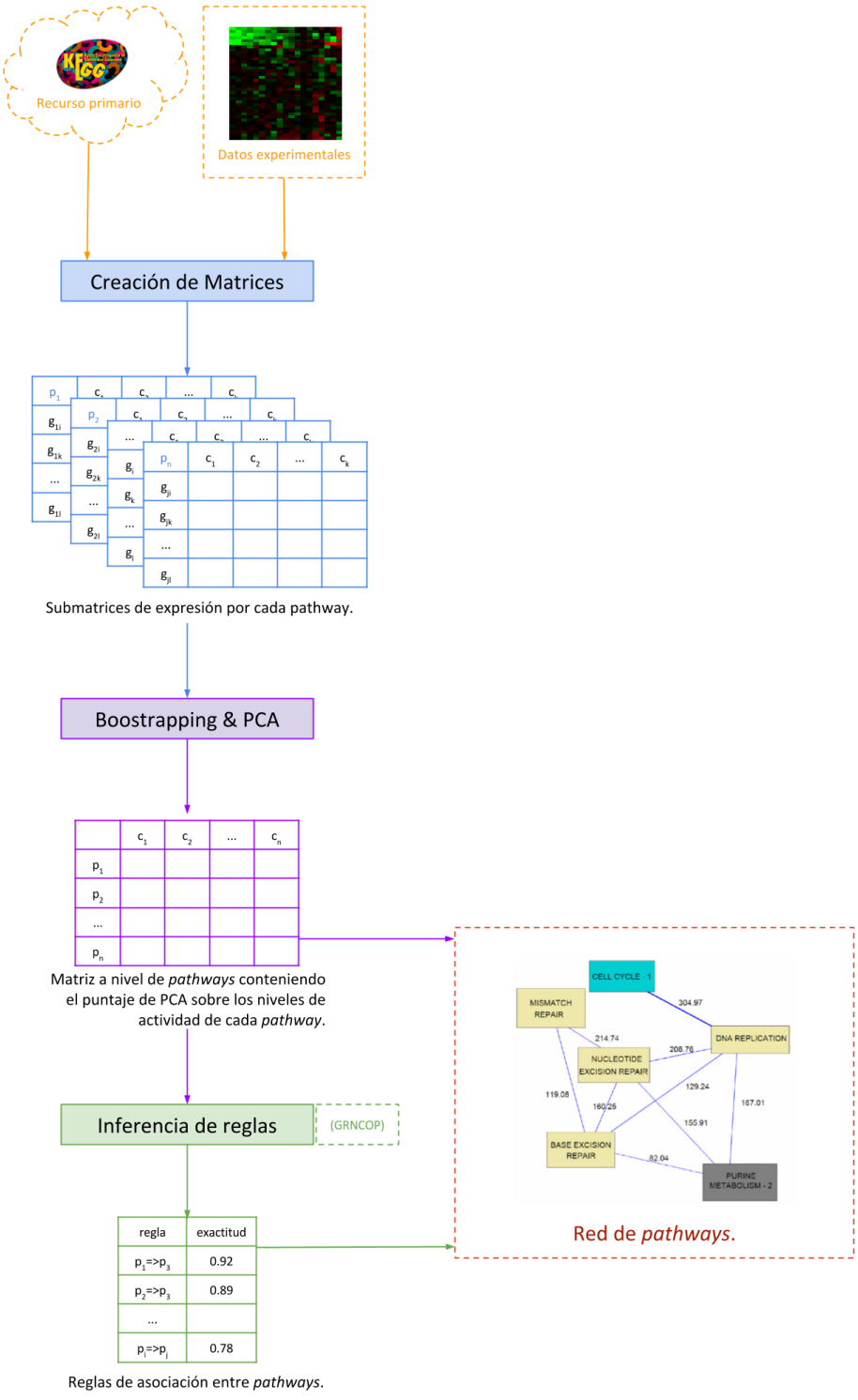

Figura 15 – Esquema del algoritmo de PANA.

Detección de redes utilizando minería de texto

Existen algoritmos que utilizan minería de textos para encontrar relaciones entre rutas biológicas y así armar una red de éstas. Estos algoritmos pretenden sobreponerse al problema del rápido crecimiento de las bases de datos de publicaciones, lo cual hace difícil la tarea de trabajar con la información actualizada. Sucede lo mismo para las redes de rutas biológicas, en cada nueva publicación puede haberse encontrado una nueva relación entre dos o más rutas biológicas o como ya se ha mencionado, los límites de las mismas no están bien definidos, por lo que es valioso poder trabajar con la última información posible.

Arizona Relation Parser

Este enfoque (Mcdonald, Chen, Su, & Marshall, 2004) se diferencia de otros mediante el uso de una gramática híbrida sintáctica-semántica. Con respecto a la parte sintáctica, para el propósito de los autores, consiste en etiquetas POS (por sus siglas en inglés, *Part Of Speech*) y alguna otra información descrita. Por otro lado la información semántica consiste en palabras y patrones de dominio específicos y se incorpora a través de una plantilla. En este método se aplican en conjunto el análisis sintáctico y el semántico,mediante el uso de un mayor número de clases de palabras, o etiquetas, que reflejan las propiedades relevantes de palabras. Esto implica que se deben escribir expresamente reglas de análisis específicas para cada clase, es decir se deben escribirmuchas reglas para apoyar las numerosas etiquetas. Sin embargo, estas reglas pueden escribirse sin léxico específico, ya que es independiente de los verbos utilizados, a diferencia de los enfoques puramente semánticos. Los resultados son volcados en un visualizador de redes.

La gran desventaja de este método es la necesidad de un experto para describir las reglas para cada clase de palabras previamente a la ejecución del algoritmo, ya que requiere mucho conocimiento para arrojar resultados significativos.

PANTex

Este método (Dussaut, Cravero, Ponzoni, Maguitman, & Cecchini, 2014) presenta un enfoque de minería de texto para ayudar a la construcción de una red de rutas biológicas. En la figura 16 se muestra un esquema de la metodología.

Como punto de partida, se utiliza KEGG (Kanehisa, Goto, Sato, Kawashima, Furumichi, & Tanabe, 2014)(Kanehisa & Goto, 2000) para reunir una lista de rutas para cada organismo, en primera instancia sólo se consideran humanos y levadura como organismos válidos. Con esta lista, utilizando *Entrez Utilities*(NCBI Resource Coordinators, 2013) y la base de datos de PubMed[1] se realiza una búsqueda para cada par de rutas contenidas en dicha lista. Estos resultados son almacenados en una matriz de intersección, llamada IRPM por sus siglas en inglés *Intersection Results Pathway Matrix*. Esta matriz cuenta con las rutas biológicas como filas y columnas y el contenido de la misma con el resultado de la búsqueda realizada. Además se realiza la búsqueda de cada ruta como término individual. Estos resultados son almacenados en el arreglo PR, por sus siglas en inglés *Pathway Results*, paraser usados en la normalización. Esta normalización se lleva a cabo calculando el índice de Jaccard:

[1]http://www.ncbi.nlm.nih.gov/pubmed

$$\frac{\left|Pw_i \cap Pw_j\right|}{\left|Pw_i \cup Pw_j\right|} = \frac{\left|Pw_i \cap Pw_j\right|}{|Pw_i| + \left|Pw_j\right| - \left|Pw_i \cap Pw_j\right|}$$

donde $|Pw_i|$ y $\left|Pw_j\right|$ son el contenido en el arreglo PR para la ruta biológica (o *pathway*) i y la ruta j respectivamente, $\left|Pw_i \cap Pw_j\right|$ es el contenido de IRPM para la ruta i y la ruta j. Esto hace que la cantidad de resultados encontrados para cada par de rutas biológicas se vea normalizado en relación a cuantas veces aparecen estas rutas en la literatura.

Para validar el método, se utilizan los datos reportados en Alexeyenko y Sonnhammer(Alexeyenko & Sonnhammer, 2009).

Los resultados de este método se vieron muy afectados y dependientes del corpus o publicaciones utilizadas.

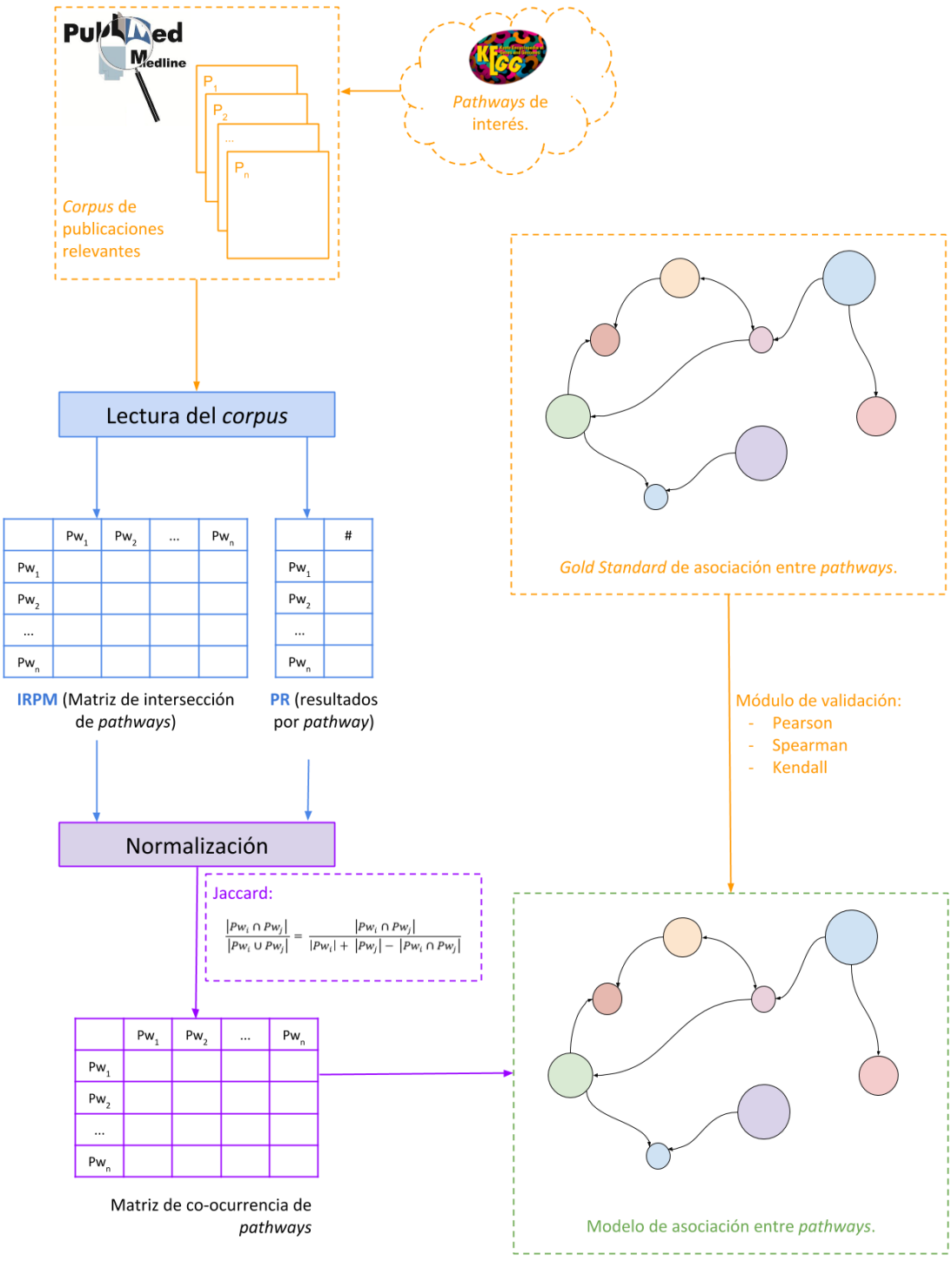

Figura 16 – Esquema del algoritmo de PANTeX.

6.3 Sumario

En este capítulo se presentaron diferentes metodologías que en la actualidad existen para inferir redes de rutas biológicas, ya sea que las utilicen para encontrarle sentido a la cantidad abrumadora de datos biológicos o que intenten encontrar interconexiones entre estas rutas. Se revisaron varios métodos computacionales relativos al estudio de interacciones entre rutas biológicas, incluyendo dos primeras contribuciones (PANA y PANTex) desarrolladas en el marco de esta tesis y en colaboración con otros grupos de investigación.

Con respecto a los métodos que detectan rutas diferencialmente expresadas, una falencia muy común en los estos métodos es que los genes dentro de la ruta son tratados por igual, independientemente de la importancia del gen dentro de ella.

Por otra parte, en la mayoría de los algoritmos que se presentaron para la inferencia de conexiones entre las rutas biológicas, se encuentra un común denominador, y es que sólo se tienen en cuenta los genes compartidos para las interacciones entre rutas, mientras que las conexiones entre genes de distintas rutas que no son compartidos entre ellas, son dejados de lado (a excepción de los métodos PathNet y PANA). Además, en todos los casos, no se incorpora directamente al proceso de inferencia la contrastación de los perfiles de expresión de los genes que están liderando la interacción entre las rutas biológicas, lo cual resulta deseable para facilitar la interpretabilidad de las interacciones. Por estas razones, entre otras, se ha propuesto un nuevo método para determinar conexiones entre rutas, el cual es presentado en el próximo capítulo de esta tesis.

CAPÍTULO 7: PET

Como ya hemos mencionado anteriormente, una ruta biológica representa una secuencia de reacciones o interacciones entre un grupo de genes expresados que participan en un proceso biológico. Durante la última década, el análisis de las rutas biológicas se ha convertido en una estrategia clave para entender el significado biológico de un grupo de genes en experimentos de alto rendimiento (*high-throughput*). Para muchos fenómenos celulares complejos es difícil explicarlos por medio de estudios que se centran sólo al nivel de genes, por esta razón, se han propuesto muchos enfoques para identificar qué rutas biológicas se enriquecen bajo alguna condición específica. Estos enfoques se han descrito en el capítulo anterior.

Sin embargo, la mayoría de los métodos existentes consideran que las rutas biológicas son entidades aisladas de una célula, sin tener en cuenta las *diafonías* entre ellas. En este contexto, la palabra "*diafonía*" (*crosstalk*) se refiere a la situación en la que uno o más componentes de una ruta afectan a otra, generando una coordinación entre los diferentes procesos biológicos. En otras palabras, los métodos fallan en la detección de las relaciones entre rutas que no están curadas y que no tienen genes en común. Para hacer frente a esta limitación, se han creado nuevos enfoques recientemente, presentados también en el capítulo anterior, que exploran el concepto de asociaciones entre rutas utilizando datos de microarrays e información topológica.

Estas estrategias detectan conexiones entre las rutas que no comparten genes en común. Sin embargo, más allá de los interesantes resultados obtenidos por estos métodos, pensamos que la sincronización entre los genes expresados diferencialmente durante el descubrimiento de asociaciones de diafonía debe ser incluido como parte del proceso de inferencia. Por lo tanto, en este capítulo se propone el uso de biclustering de datos de expresión de genes, en combinación con el análisis topológico, con el fin de extraer las asociaciones sincronizadas de rutas biológicas. De este modo esperamos obtener un método que facilite la interpretabilidad de la relevancia biológica de las interacciones encontradas entre las rutas biológicas.

El nuevo método fue bautizado PET, por sus siglas en inglés *Cross-talk **P**athway Inference by using Gene **E**xpression Data Biclustering and **T**opological Information*, y busca inferirredes de rutas biológicas basándose en su topología y ayudándose con los datos de expresión de genes de un experimento de dos clases (p.e. control vs. afectados) mediante un análisis de biclustering, logra encontrar sentido a los resultados del experimento con la suficiente información biológica.

En la siguiente sección describiremos el método PET. Después presentaremos una evaluación de su desempeño contrastándolo con los métodos PathNet y PANA mediante el uso de datos de expresión de genes relacionados con la enfermedad de Alzheimer. Finalmente, discutiremos las principales conclusiones de esta contribución.

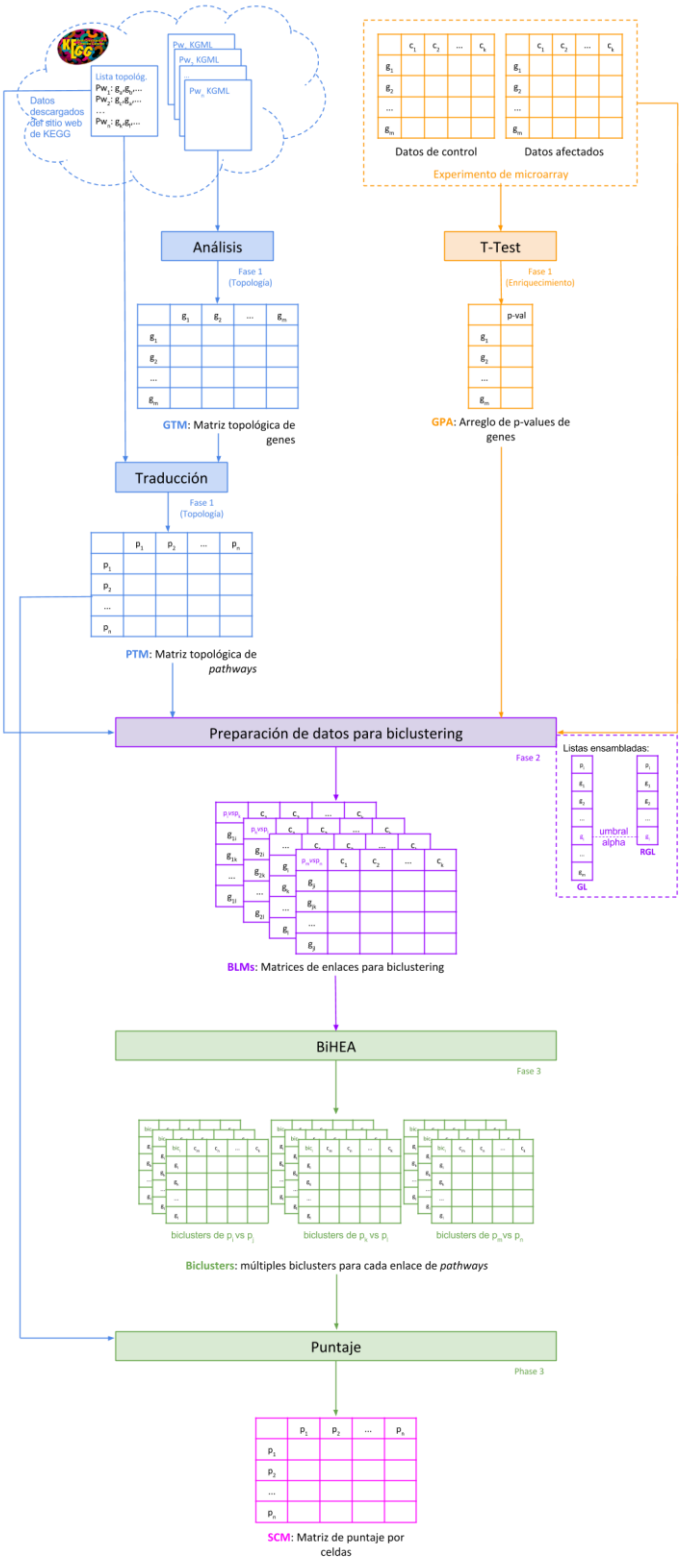

Figura 17 – Esquema del algoritmo de PET.

7.1 Algoritmo

El algoritmo puede ser desglosado en tres fases. En la primera se produce el armado de las matrices y arreglos necesarios a partir de los datos de entrada; estos datos de entrada serán la información topológica de KEGG (Kanehisa, Goto, Sato, Kawashima, Furumichi, & Tanabe, 2014)(Kanehisa & Goto, 2000) y dos matrices resultantes de un experimento de microarray del tipo de control vs. afectados. Luego, en la segunda fase, estas matrices son utilizadas para el ensamblado de nuevos datos para la posterior ejecución del algoritmo de biclustering. Por último, en la tercera fase, se procede a ejecutar el algoritmo de biclustering BiHEA (Gallo, Carballido, & Ponzoni, BiHEA: A Hybrid Evolutionary Approach for Microarray Biclustering, 2009), mediante la utilización de la herramienta BAT(Gallo, Dussaut, Carballido, & Ponzoni, 2010), y con los datos arrojados por la misma se calcula el puntaje final para cada posible relación de rutas biológicas, otorgando así información necesaria para ensamblar una red de rutas biológicas para su posterior análisis. Un esquema del algoritmo puede apreciarse en la figura 17.

Fase 1: Armado de matrices y arreglos

En esta fase se ensamblan las matrices de información topológica y el arreglo del análisis de enriquecimiento a partir de datos de expresión genética.

En primera instancia se necesitan los resultados de un experimento de microarray de dos clases (control vs. afectados). Con estos datos se realiza el clásico t-test para encontrar los

p-values, es decir los genes que están diferencialmente expresados en las muestras. Estos datos son almacenados en un arreglo llamado GPA (*gene p-value array*).

Construimos también una matriz similar al *pooled pathway* de PathNet (Dutta, Wallqvist, & Reifman, 2012), que se denomina GTM (*gene topology matrix*). A diferencia de a la matriz de PathNet, en GTM se utilizantodos los genes de todas las rutas biológicas que se encuentran en la base de datos de KEGG, para de esta forma realizar de forma más amplia, pruebas sobre ella. Esto se realiza descargando todos los archivos KGML provistos por KEGG y ensamblando la matriz de la siguiente forma: en la fila i y la columna j el contenido de GTM será:

- 1 si el gen g_i y el gen g_j tienen una conexión en al menos una ruta biológica.
- 0 en caso contrario.

Esta matriz GTM que es la representación de una red de genes total, se la mapea a otra matriz, denominada PTM (*pathway topology matrix*), en donde se encuentra ahora la red de rutas biológicas. Como antes se ha mencionado, los enlaces de estas rutas biológicas se calculan para aquellas que posean genes en común y aquellas que posean enlaces entre sus genes pertenecientes. Es decir, el contenido de PTM en la fila i y la columna j es:

- 1 si la ruta biológica p_i y la ruta biológica p_j tienen genes en común (figura18 A y figura18 D).
- 1 si la ruta biológica p_i y la ruta biológica p_j no tienen genes en común pero uno o más genes de p_i tiene conexiones con uno o más genes de p_j (figura18 B y figura18 E).
- 0 en caso contrario. Es decir las rutas biológicas p_i y p_j no tienen genes en común y no existe ninguna conexión entre sus genes (figura18 C y figura18 F).

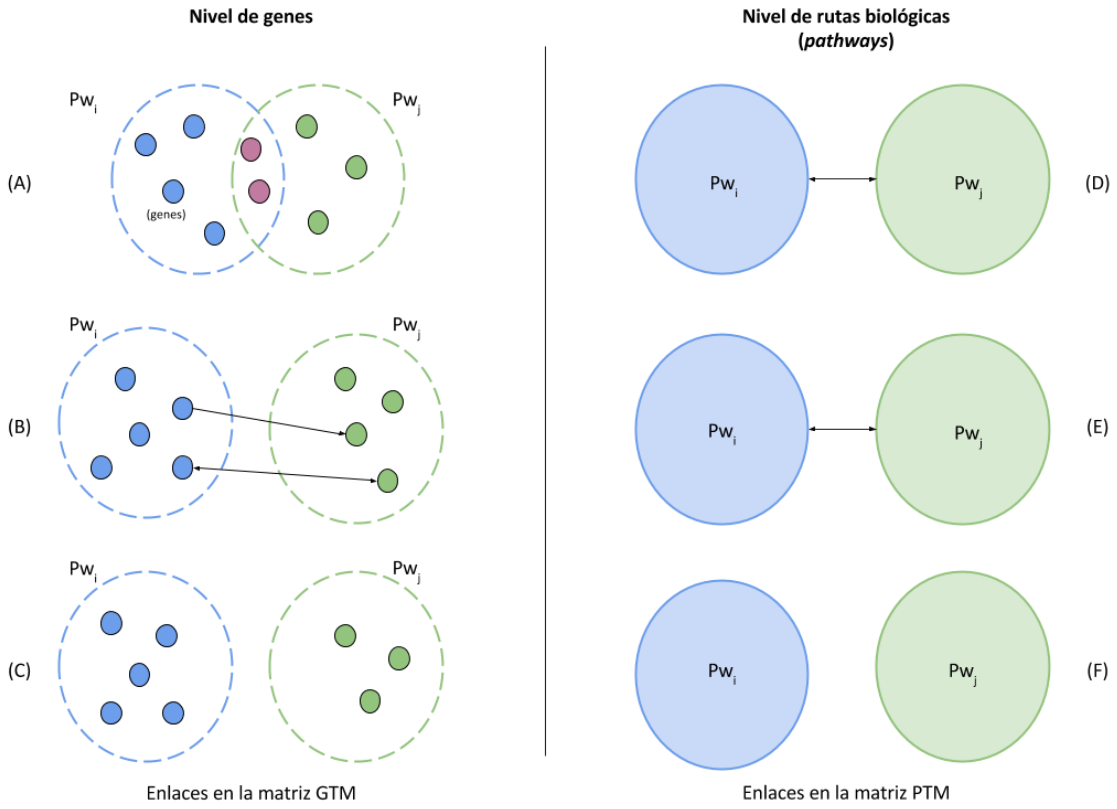

Figura 18 - Relaciones topológicas entre dos rutas biológicas (Pw_i y Pw_j). En la (A) ambas rutas tienen genes en común y, en (D) esta situación se refleja en la matriz PTM con un vínculo entre esas rutas. En (B) existe una conexión topológica entre los genes de las diferentes rutas, y en (E) esta situación se refleja también con un enlace en PTM. En (C) no hay evidencia de una relación topológica entre los genes de ambas rutas, por lo tanto, en (F) no hay conexión entre las dos rutas.

Fase 2: Preparación de los datos de biclustering

Por cada enlace o 1 en PTM, es decir por cada par de rutas biológicas (p_i, p_j) que tienen una conexión o enlace, se arma una nueva matriz, llamada BLM por sus siglas en inglés *Biclustering Link Matrix*, de la siguiente forma:

1) Se obtiene una lista de los genes de ambas rutas biológicas (GL) y de esa lista se obtienen sólo los relevantes o los que se encuentran enriquecidos por los datos de microarray. Estos genes relevantes se vuelcan en una nueva lista (RGL) (figura 19 A y figura 19 B). La determinación si un gen está o no enriquecido se hace mediante su valor en el arreglo GPA, sólo si este valor es menor a un umbral llamado *alpha* establecido en 0.05, entonces se dice que el gen está enriquecido para poder formar parte de la lista RGL.
2) Con los genes de la lista RGL se realiza un control antes de armar la matriz BLM (figura 19 C y figura 19 D):
    - Si los genes de RGL pertenecen solamente a uno de las 2 rutas biológicas del par, entonces no se realiza la matriz BLM y se dice que el enlace entre p_i y p_j no se encuentra expresado en los datos del experimento de microarray, por lo que se le otorga un "puntaje" de 0, sin siquiera pasar por el algoritmo de biclustering.
    - Si en cambio, RGL contiene genes de ambas rutas biológicas, podemos hablar de un enlace entre rutas enriquecido y procedemos a ensamblar la matriz BLM. Esta matriz es simplemente una submatriz de los datos de microarray afectados, conteniendo sólo los genes de RGL como filas.

Figura 19 – Ejemplo de relaciones topológicas entre los genes de varias rutas biológicas (A y B) y su traducción a relaciones entre rutas (CyD). En (A) las relaciones son meramente topológicas y en(B) las relaciones son filtradas por los genes expresados. (C) es la traducción al mundo de las rutas biológicas de (A) y (D) es la traducción de (B), en (D) se puede ver que hay rutas o enlaces que no son relevantes para el experimento dado; este es el resultado de la falta de expresión de sus genes en el análisis de enriquecimiento.

Fase 3: Algoritmo de biclustering y puntuación.

El último paso consiste en la ejecución de un algoritmo biclustering, BiHEA (Gallo, Carballido, & Ponzoni, 2009)a través de la herramienta BAT (Gallo, Dussaut, Carballido, & Ponzoni, 2010)para cada matriz BLM montado en Fase 2, con el fin de obtener un conjunto de biclusters para los posibles vínculos entre las rutas. Solamente se consideran aquellos biclusters que contienen al menos un gen de cada ruta biológica presente en la asociación. Entonces, una puntuación para cada enlace se calcula contando las celdas que pertenecen a estos biclusters. Los biclusters encontrados por BiHEA no sólo dan una puntuación al enlace, sino que también explican la conexión entre las rutas, dando la oportunidad de analizar los patrones de co-expresión y el comportamiento de los genes involucrados.

7.2 Resultados

Se utilizaron dos conjuntos de datos de microarrays, que se generaron para el estudio de la enfermedad de Alzheimer (AD por sus siglas en inglés *Alzheimer Disease*). Ambos conjuntos de datos fueron descargados de la base de datos de Gene Expression Omnibus (GEO[2]) y se utilizaron previamente por Dutta et al.(Dutta, Wallqvist, & Reifman, 2012) y Ponzoni et al. (Ponzoni, y otros, 2014) para la detección de las rutas biológicas asociadas con la AD mediante los métodos PathNet y PANA respectivamente.

[2]http://www.ncbi.nlm.nih.gov/geo/

El primer conjunto de datos (GEO ID: GSE1297) se empleó en el estudio del perfil de expresión de genes de la región del hipocampo del cerebro como una función de la progresión de la enfermedad (incipiente, moderada y grave). El segundo conjunto de datos (GEO ID: GSE5281) se usó para explorar el efecto de la AD en seis regiones cerebrales diferentes: la corteza entorrinal, campo del hipocampo CA1, circunvolución temporal media, la corteza cingulada posterior, gyrus frontal superior y la corteza visual primaria. Dutta et al (Dutta, Wallqvist, & Reifman, 2012) y Ponzoni et al. (Ponzoni, y otros, 2014) analizaron las rutas que tienen asociación estadísticamente significativa con la ruta de AD (KEGG id hsa05010). El análisis fue realizado sobreseis condiciones: muestras moderadas (MOD) y severas (SEV) en el conjunto de datos progresión de la enfermedad; y la corteza visual primaria (VCX), campo del hipocampo CA1 (HIP), la circunvolución temporal media (MTG) y regiones corteza cingulada posterior (PC) en el conjunto de datos regiones del cerebro. Solo las reglas o rutas biológicas que se encuentran en al menos 3 condiciones fueron seleccionadas como asociaciones pertinentes en ambos estudios.

Siguiendo este mismo diseño experimental, se ejecutó el método PET para los mismos 6 conjuntos de muestras y se encontraron, como se esperaba, que las reglas correspondientes a las diafonías entre rutas biológicas relacionadas con la AD se acumulan en los percentiles más altos (figura 20). Seeligió el primer percentil de cada experimento como resultado las reglas asociadas a la AD, es decir se tomó solamente el 5% del universo de reglas, el 5% con mayor puntaje. Estas reglas se clasifican en un orden descendente según la puntuación del método. Tomando esto en consideración, 63 reglas están relacionadas con AD (es decir, las rutas están vinculadas a la ruta hsa05010) con una frecuencia de ocurrencia entre los experimentos o condiciones de 3 o más. Dado que, también se obtiene un conjunto de biclusters para cada asociación en cada experimento, y es posible calcular un *bicluster de consenso*, que representa el conjunto máximo de genes

que se comportan de manera similar en la mayoría de los experimentos. Este bicluster proporciona información sobre el significado de la asociación.

Estas reglas se comparan con las 36 encontradas por PANA y PathNet. De las 36 reglas, PET encuentra 29: 2 de las 7 reglas restantes (hsa04960 y hsa05218) son encontradas por el método, pero con una menor frecuencia, otras 2 (hsa00620 y hsa04512) no están topológicamente vinculadas con la ruta biológica de AD, y las 3 restantes (hsa04320, hsa05213 y hsa04973) no se encuentran en el primer percentil elegido anteriormente.

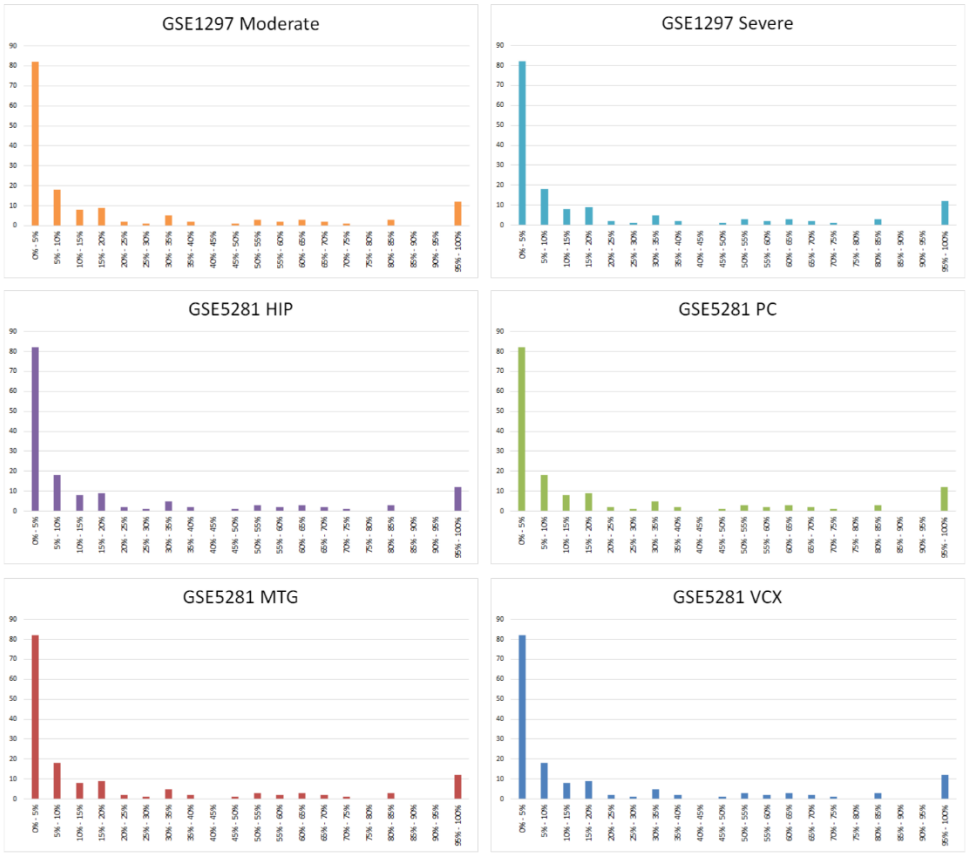

Figura 20 – Cantidad de rutas relacionadas con AD en los percentiles de todas las reglas inferidas en cada condición: GSE1297 moderado (*Moderate*), GSE1297 severo (*Severe*), GSE5281 hipocampo CA1(*HIP*), GSE5281 corteza posterior cingulada (*PC*), GSE5281 Medio Temporal Gyrus (*MTG*) y corteza visual primaria (*VCX*).

La tabla 1 muestra las 63 reglas resultantes relacionadas con AD. Las dos primeras columnas muestran el nombre de la ruta vinculada a AD y la frecuencia tomando solo el primer percentil de las 6 condiciones. Las columnas 3 y 4 muestran si la regla fue encontrada por otro enfoque y la correspondiente frecuencia informada, respectivamente. Las columnas 5 y 6 representan los genes que forman el *bicluster consenso* y la frecuencia correspondiente, respectivamente. Por último, la 7º y 8º columnas muestran el término de enriquecimiento y el p-value obtenido con las herramientas de DAVID (Huang, Sherman, & Lempicki, 2009)(Huang, Sherman, & Lempicki, 2009) por el conjunto de genes del *bicluster de consenso*. Esto demuestra que la mayoría de las asociaciones de la ruta reportados en esta tabla contienen *biclusters de consenso* que, además de la alta co-expresión de sus perfiles de genes, también tienen similitud semántica.

Tabla 1 - Tabla de diafonía entre rutas. Asociacionesencontradas por el algoritmo de PET.Las dos primeras columnas muestran el nombre de la ruta vinculada a AD y la frecuencia con respecto a las 6 condiciones o experimentos realizados. Las columnas 3 y 4 muestran si la regla fue encontrada o no por otro enfoque y la frecuencia correspondiente. Las columnas 5 y 6 representan los genes en el bicluster de consenso encontrado para esa asociación y la frecuencia correspondiente. Por último, la 7º y 8 º columnas muestran el término enriquecimiento y p-value obtenido con las herramientas de DAVID para los biclusters de consenso.

Rule	Frequency	Other Methods	Max Frequency of Other Methods	Genes in the Consensus Bicluster	Frequency of the Consensus Bicluster	Enrichment Term of the Consensus Bicluster	P value of the Enrichment
hsa04070	6	PathNet	3	ATP2A2 (hsa05010), PI4KA (hsa04070)	6	SP_PIR_KEYWORDS atp-binding	0.068936834
hsa00190	6			ATP2A2 (hsa05010), ATP6V1E1 (hsa00190)	6	GO:0019829 cation-transporting ATPase activity	0.002772857
hsa04145	6			ATP2A2 (hsa05010), ATP6V1E1 (hsa04145)	6	GO:0019829 cation-transporting ATPase activity	0.002772857
hsa05110	6			ATP2A2 (hsa05010), ATP6V1E1 (hsa05110)	6	GO:0019829 cation-transporting ATPase activity	0.002772857
hsa05120	6			ATP2A2 (hsa05010), ATP6V1E1 (hsa05120)	6	GO:0019829 cation-transporting ATPase activity	0.002772857
hsa01100	6			UQCRC2 (hsa05010 and hsa01100), MDH2 (hsa01100)	6	GO:0009060 aerobic respiration	0.002587226
hsa04360	6	Both	4	ATP2A2 (hsa05010), PPP3CA (hsa05010 and hsa04360)	6	GO:0006816 calcium ion transport	0.010496747
hsa04020	6			ATP2A2 (hsa05010, hsa04020), PPP3CA (hsa05010, hsa04020)	6	GO:0006816 calcium ion transport	0.010496747
hsa05200	6	Both	4	PPP3CA (hsa05010), UQCRC2 (hsa05010)	6	GO:0006796 phosphate metabolic process	0.071924897
hsa04114	6	Both	4	NDUFA8 (hsa05010), PPP2CA (hsa04114)	6	GO:0006793 phosphorus metabolic process	0.071924897
hsa04120	6	PathNet	4	UQCRC2 (hsa05010), UBE2N (hsa04120)	6	GO:0006508 proteolysis	0.077912478
hsa04810	6	PathNet	6	ATP2A2 (hsa05010), PFN2 (hsa04810)	6	-	-
hsa05012	6	PathNet	5	SLC25A5 (hsa05012), ATP5O (hsa05012 and hsa05010)	5	SP_PIR_KEYWORDS mitochondrion inner membrane	0.010033793
hsa03050	6	PathNet	3	ATP5F1 (hsa05010), PSMA1 (hsa03050)	5	SP_PIR_KEYWORDS hydrolase	0.080842215

hsa04012	6	PathNet	3	ATP2A2 (hsa05010), MAP2K4 (hsa04012)	5	SP_PIR_KEYWORDS ATP	0.012269301
hsa04062	6			ATP2A2 (hsa05010), PPP3CA (hsa05010)	5	GO:0006816 calcium ion transport	0.010496747
hsa05016	6	PathNet	5	ATP5F1 (hsa05016 and hsa05010), ATP5O (hsa05016 and hsa05010)	5	GO:0000276 mitochondrial proton-transporting ATP synthase complex, coupling factor F(o)	7.04E-04
hsa04141	6			ATP5F1 (hsa05010), ATP5O (hsa05010)	5	GO:0000276 mitochondrial proton-transporting ATP synthase complex, coupling factor F(o)	7.04E-04
hsa04910	6	Both	4	SNCA (hsa05010), PRKCZ (hsa04910)	4	GO:0050806 positive regulation of synaptic transmission	0.002513306
hsa04666	6	PathNet	3	NDUFS3 (hsa05010), RAC1 (hsa04666)	4	GO:0032535 regulation of cellular component size	0.020032525
hsa04310	6			SNCA (hsa05010), RAC1 (hsa04310)	4	GO:0030100 regulation of endocytosis	0.004509166
hsa05100	6	PathNet	6	SNCA (hsa05010), WASL (hsa05100)	4	GO:0009617 response to bacterium	0.014266706
hsa04010	6	Both	3	NDUFAB1 (hsa05010), RASGRF2 (hsa04010)	4	GO:0005509 calcium ion binding	0.070784873
hsa04144	6	PathNet	4	NDUFS3 (hsa05010), TSG101 (hsa04144)	4	GO:0001558 regulation of cell growth	0.014340627
hsa04110	6			UQCR11(hsa05010), CCND3 (hsa04110)	2	-	-
hsa00230	6			-	-	-	-
hsa04270	5			CDK5 (hsa05010), GUCY1B3 (hsa04270)	5	SP_PIR_KEYWORDS nucleotide-binding	0.087652716
hsa04540	5	Both	5	ATP2A2 (hsa05010), PPP3CA (hsa05010), GUCY1B3 (hsa04540)	5	SP_PIR_KEYWORDS metal-binding	0.023866551
hsa04972	5			NDUFA8 (hsa05010), ATP2A2 (hsa05010 and hsa04972)	5	GO:0031090 organelle membrane	0.08574558
hsa04914	5	PathNet	3	ATP2A2 (hsa05010), PPP3CA (hsa05010)	5	GO:0006816 calcium ion transport	0.010496747
hsa04650	5			ATP2A2 (hsa05010), PPP3CA (hsa05010 and hsa04650)	5	GO:0006816 calcium ion transport	0.010496747
hsa04916	5			COX6C (hsa05010), PRKCB (hsa04916)	5	-	-
hsa04912	5	Both	5	ATP5A1 (hsa05010), MAP2K4 (hsa04912)	4	UP_SEQ_FEATURE nucleotide phosphate-binding region:ATP	0.050332235
hsa00010	5			ATP5A1 (hsa05010), ATP5B (hsa05010)	4	IPR000793: ATPase, F1/V1/A1 complex, alpha/beta subunit, C-terminal	3.00E-04
hsa00020	5	PANA	4	ATP5A1 (hsa05010), MDH1 (hsa00020)	4	GO:0006091 generation of precursor metabolites and energy	0.023137197
hsa03013	5			ATP5J (hsa05010), STRAP (hsa03013)	4	-	-
hsa04080	5			COX5B (hsa05010), NDUFA4 (hsa05010)	3	SP_PIR_KEYWORDS respiratory chain	0.003691188
hsa04660	5			NDUFA4 (hsa05010), MAP2K2 (hsa04660)	3	GO:0016310 phosphorylation	0.059136606
hsa04510	5	Both	5	NDUFAB1 (hsa05010), ATP5J (hsa05010), RAC1 (hsa04510)	3	GO:0006119 oxidative phosphorylation	0.014436521
hsa04530	5			NDUFB2 (hsa05010), CTNNA2 (hsa04530)	3	-	-
hsa04971	4			CDK5 (hsa05010), PRKCB (hsa04971)	4	SP_PIR_KEYWORDS serine/threonine-specific protein kinase	0.003119314
hsa04730	4	Both	4	ATP2A2 (hsa05010), PPP3CA (hsa05010), GUCY1B3 (hsa04730)	4	SP_PIR_KEYWORDS metal-binding	0.023866551
hsa03015	4			PPP3CB (hsa05010), PPP2CA (hsa03015)	4	IPR006186:Serine/threonine-specific protein phosphatase and bis(5-nucleosyl)-tetraphosphatase	7.80E-04
hsa05131	4	PathNet	4	ATP2A2 (hsa05010), ATP5C1 (hsa05010)	4	GO:0019829 cation-transporting ATPase activity	0.002772857
hsa05323	4			ATP2A2 (hsa05010), ATP6V1E1 (hsa05323)	4	GO:0019829 cation-transporting ATPase activity	0.002772857
hsa05160	4			NDUFB2 (hsa05010), MAPK9 (hsa05160)	4	GO:0016310 phosphorylation	0.059136606
hsa00330	4			PPP3CA (hsa05010), GOT2 (hsa00330)	4	GO:0010033 response to organic substance	0.053296866
hsa04380	4			ATP2A2 (hsa05010), PPP3CA (hsa05010 and hsa04380)	4	GO:0006816 calcium ion transport	0.010496747
hsa04970	4			PPP3CA (hsa05010), PRKCB (hsa04970)	4	GO:0006816 calcium ion transport	0.010496747
hsa04722	4	Both	4	MAGED1 (hsa04722), PPP3CA (hsa05010)	4	-	-
hsa04370	4	PANA	3	COX6C (hsa05010), PRKCB (hsa04370)	4	-	-
hsa04664	4			COX6C (hsa05010), PRKCB (hsa04664)	4	-	-
hsa05142	4			ATP5B (hsa05010), ATP5O (hsa05010)	3	GO:0045261 proton-transporting ATP synthase complex, catalytic core F(1)	4.69E-04
hsa04520	4	Both	5	ATP5F1 (hsa05010), ATP5O (hsa05010)	3	GO:0000276 mitochondrial proton-transporting ATP synthase complex, coupling factor F(o)	7.04E-04

hsa04962	4			DYNC1I1 (hsa04962), COX7A2L (hsa05010)	3	-	-
hsa05211	3			NDUFA8 (hsa05010), MAP2K1 (hsa05211)	3	GO:0016310 phosphorylation	0.059136606
hsa04720	3	Both	5	PPP3CA (hsa04720 and hsa05010), ATP2A2 (hsa05010)	3	GO:0006816 calcium ion transport	0.010496747
hsa04146	3	PANA	3	COX7B (hsa05010), PEX11B (hsa04146 and hsa05010)	3	GO:0005739 mitochondrion	0.085041465
hsa04260	3			NDUFV2 (hsa05010), SLC9A6 (hsa04260)	3	GO:0005739 mitochondrion	0.085041465
hsa05214	3	PathNet	3	CDK5 (hsa05010), MAP2K1 (hsa05214)	3	BIOCARTA h_cdk5 Pathway: Phosphorylation of MEK1 by cdk5/p35 down regulates the MAP kinase pathway	0.008350731
hsa05130	3	PathNet	5	ARPC3 (hsa05130), UQCRC1 (hsa05010)	3	-	-
hsa04662	3			COX6C (hsa05010), PRKCB (hsa04662)	3	-	-
hsa05146	3			COX6C (hsa05010), PRKCB (hsa05146)	3	-	-

7.3 Relevancia biológica de los resultados

En las próximas secciones, dos ejemplos de análisis se presentan con el fin de ilustrar cómo utilizar los biclusters de consenso en la interpretación biológica de las diafonías. En el primer caso, se seleccionó una asociación de rutas detectada por PathNet, PANA y PET, y para esta regla se analiza la sincronización entre los perfiles de los genes de consenso. En el segundo caso, se seleccionó una asociación de rutas informada exclusivamente por PET y se discuten los patrones de expresión comunes entre los genes de consenso y los genes expresados diferencialmente con mayor frecuencia.

Diafonía entre las rutas de AD y unión GAP

La asociación entre las rutas de AD y GAP es inferida por los tres métodos: PathNet, PANA y PET. Además, varias referencias en la literatura apoyan esta relación con su explicación biológica(Cruz, Ball, & Dienel, 2010)(Mei, Ezan, Giaume, & Koulakoff, 2010).

En particular, es posible analizar con PET los genes de consenso, que son ATP2A2 (ruta AD), PPP3CA (ruta AD) y GUCY1B3 (ruta GAP). El bicluster de consenso integrado por estos genes se encuentra en 5 condiciones. La figura 21 muestra los perfiles de expresión de estos genes a lo largo de estas condiciones, donde es clara la sincronización entre estos genes en todos los casos.

Figura 21 – Gráficos de perfiles de expresión de los genes contenidos en el bicluster de consenso entre la ruta de AD (hsa05010) y la ruta de unión GAP(hsa04540). Este bicluster de consenso se encuentra en 5 de las 6 condiciones.

En cuanto el significado biológico de estos genes, ATP2A2 codifica uno de los SERCA Ca(2+)-ATPases, que son bombas intracelulares situadas en la retícula sarcoplásmica o endoplasmática de las células musculares. Esta enzima cataliza la hidrólisis de ATP junto con la translocación de calcio desde el citosol al lumen de la retícula sarcoplásmica, y está implicado en la sustracción de calcio asociado con la excitación muscular y la contracción. Estas disfunciones están implicadas en la progresión del cáncer y varias enfermedades neurológicas(Brini & Carafoli , 2009).

PPP3CA, una fosfatasa Ser/Thr calmodulina-dependiente, también conocida como calcineurina A alfa, está relacionada con varias actividades biológicas, actuando como un modificador Ca(2+)-dependiente de estado de fosforilación. En particular, este gen se asocia con la regulación de la transmisión sináptica. Nuevas isoformas empalmadas de forma alternativa de PPP3CA fueron recientemente identificadas y sus expresiones fueron encontradas alteradas en regiones cerebrales de la enfermedad después de la muerte de pacientes de Alzheimer (Chiocco, y otros, 2010). Estos datos demuestran la importancia del gen PPP3CA en el mecanismo molecular relacionado con esta enfermedad.

Orsetti et al. (Orsetti, Di Brisco, Canonico, Genazzani, & Ghi, 2008) estudian la regulación de genes en la corteza frontal de ratas expuestas al paradigma de estrés leve crónico. En particular, estudian la expresión diferencial de genes asociados con la transmisión sináptica y transporte de iones de metal en ratas anhedónicas. Muestran que las transcripciones que caen en la categoría de transporte de iones de metal, todos subexpresadas, están relacionados principalmente con el transporte de Ca2+ e incluyen genes que codifican las a-isoformas de la calcineurina A y B (PPP3R1 y PPP3CA), la a-isoforma de Ca2+ / proteína quinasa II calmodulina-dependiente (CAMK2A), y la bomba Ca2+ SERCA2 (ATP2A2). Por lo tanto se desprende de los experimentos de Orsetti la existencia de una fuerte asociación

entre los comportamientos de PPP3CA y ATP2A2 en el contexto de las enfermedades neurológicas.

Por otro lado, el gen GUCY1B3 codifica la subunidad soluble beta-1 guanilato ciclasa en los seres humanos, que es una proteína heterodimérica que cataliza la conversión de GTP al segundo mensajero cGMP y es parte de la ruta de señalización de la proteína quinasa óxido nítrico (NO)/sGC/cGMP-dependiente (PKG) que juega un papel clave en el procesamiento de la memoria (Shen, Li, Shou, & Cui, 2012)(Bartus, Pigott, & Garthwaite, 2013). La inhibición de GCs, PKG o fosfodiesterasa cGMP-degradantes se ha encontrado (Monfort, y otros, 2004) que afecta la potenciación a largo plazo (LTP por sus siglas en inglés *Long Time Potentiation*). En particular, Uddin y Singh (Uddin & Singh, 2013), estudiaron el envejecimiento y deterioro del aprendizaje espacial asociado a la edad (ASLI) en el hipocampo en modelos animales, encontraron que GUCY1B3 se había subexpresadas en los animales de edad avanzada, lo que puede explicar la ASLI en estos animales.

En resumen, PPP3CA y GUCY1B3 están relacionados con la potenciación a largo plazo (LTP), mientras que la relación entre PPP3CA y ATP2A2 es reportada por los estudios de Orsetti sobre la regulación de genes en la corteza frontal. Por lo tanto, la sincronización entre GUCY1B3 y los genes ATP2A2 y PPP3CA de la ruta biológica de AD es biológicamente coherente.

Diafonía entre las rutas de AD y Señalización de Calcio

El calcio es una molécula de señalización muy importante. El ion juega un papel fundamental en la señalización neuronal. Las anomalías en la regulación del calcio han sido reportadas en varias enfermedades neurodegenerativas por numerosos estudios (LaFerla,

2002)(Mattson & Chana, 2003)(Marambaud, Dreses-Werringloer, & Vingtdeux, 2009)(Hermes, Eichhoff, & Garaschuk, 2010). En particular, se ha subrayado la importancia de procesos celulares dependientes de calcio en la caracterización bioquímica de AD, lo que sugiere que las anormalidades en la homeostasis del calcio (Ca2+) podrían estar implicadas en la fisiopatología de la enfermedad (Zündorf & Reiser, 2011).

La figura 22 representa los genes de las rutas de AD y de señalización de calcio con expresión diferencial en todas las condiciones. Los genes ATP2A2 y PPP3CA, en púrpura, son compartidos por ambas rutas e integran un fuerte bicluster de consenso presente en las 6 condiciones. Sus relaciones biológicas con el AD se han explicado anteriormente. Los genes NDUFA8, UQCRC2, COX7B y ATP5C1, en azul, pertenecen a la ruta de AD y están relacionados con la regulación de las mitocondrias. En este sentido, Brzyska y Elbaum (Brzyska & Elbaum, 2003) discutieron la importancia de las mitocondrias como sitio regulador involucrado en la patogénesis de la degeneración neuronal en relación con las alteraciones del Ca2+ durante la progresión de AD. Finalmente, el gen PRKCB, en verde, sólo aparece en la ruta de señalización de calcio. Este gen tiene un papel central en la regulación de la autofagia, que es el principal sistema intracelular de degradación y constituye un mecanismo esencial en diversos eventos biológicos. Las observaciones recientes indican que la autofagia es modulada en respuesta a la situación de la energía del compartimento mitocondrial, lo que explica la sincronización entre PRKCB y los genes reguladores mitocondriales. En particular, estudios recientes (Patergnani, y otros, 2013) revelan que la modificación de los niveles de autofagia puede ser útil en la lucha contra los trastornos neurodegenerativos, que se caracterizan por niveles reducidos de autofagia.

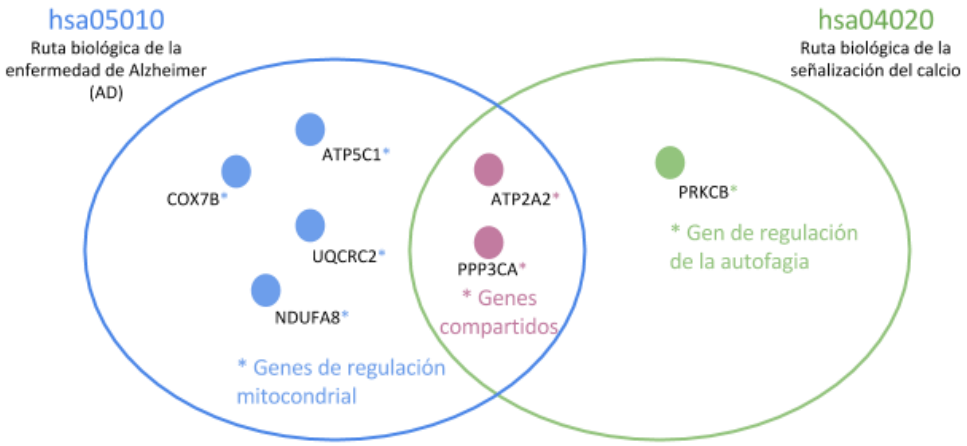

Figura 22 – Genes diferencialmente expresados asociados a la diafonía entre la ruta de AD (hsa05010) y la ruta de señalización de calcio (hsa04020). Los genes compartidos por ambas rutas están en rosa. Los genes que pertenecen a la ruta de AD o a la ruta de señalización de calcio exclusivamente, están en azul y en verde, respectivamente.

Finalmente, en la figura 23, se contrastan los perfiles de expresión de estos 7 genes. Una coordinación entre estos genes es evidente en la mayoría de las condiciones. Sin embargo, incluso con la clara expresión diferencial existente en estos siete genes en todos los conjuntos de datos y el fuerte bicluster de consenso integrado por ATP2A2 y PPP3CA, esta asociación de rutas no logra ser detectada por los métodos PathNet y PANA.

Figura 23 – Gráficos de perfiles de expresión de los genes diferencialmente expresados en la diafonía entre la ruta de AD (hsa05010) y la ruta de señalización de calcio (hsa04020). Las líneas rosadas pertenecen a los genes contenidos en el bicluster de consenso, y las líneas azules y verdes son los genes diferenciales expresados de esta diafonía (ver figura 22) a lo largo de las muestras del bicluster de consenso.

En resumen, las asociaciones de las rutas inferidas por PET son consistentes con los conocimientos previos sobre la enfermedad y logra una cobertura significativa del conjunto de asociaciones detectadas por PathNet y PANA.

Por otro lado, para todas las asociaciones inferidas por PET, un análisis ontológico muestra que los biclusters con mayor consenso a lo largo de los conjuntos de datos de microarrays están integrados por los genes relacionados, que tienen una relación semántica bien establecida en términos de anotaciones funcionales.

7.4 Sumario

La comprensión de los fenómenos biológicos que se producen a nivel celular requiere desentrañar los complejos mecanismos de interacción entre los diferentes procesos celulares. Una estrategia para hacer frente a esta tarea es identificar las *diafonías* (*crosstalks*) entre las rutas biológicas. Esto implica la detección de mecanismos de señalización entre los procesos biológicos de la coordinación que tiene lugar entre los genes que subyacen en las diferentes rutas.

Se han propuesto varios enfoques para la inferencia de diafonías a partir de los datos de expresión génica. En este capítulo, se presentó un método que utiliza la combinación de una herramienta de análisis de enriquecimiento basado en la información topológica proporcionada por KEGG y biclustering de datos de expresión génica. En pocas palabras, dado un conjunto de datos de microarrays y un par de rutas en estudio, el análisis de enriquecimiento identifica qué partes de las rutas son co-activadas por los genes con expresión diferencial en los conjuntos de datos, mientras que el análisis biclustering captura la sincronización entre los perfiles de expresión de estos genes.

El rendimiento de esta nueva metodología, llamada **PET** (por sus siglas en inglés *Crosstalk **P**athway Inference by using Gene **E**xpression Data Biclustering and **T**opological Information*), fue contrastado con dos métodos actuales arrojando resultados positivos

tanto con las diafonías ya encontradas por estos métodos, como con diafonías descubiertas por PET y posteriormente analizadas en detalle. En particular, se presenta el análisis detallado de las dos asociaciones entre la ruta de AD con la ruta GAP y con la ruta de señalización de calcio, donde se ilustra la importancia biológica de la información extraída por el biclustering.

CAPÍTULO 8: Conclusiones

El objetivo general de esta tesis consistió en diseñar nuevas técnicas computacionales para asistir a expertos en bioinformática en la obtención de nuevos conocimientos sobre el funcionamiento de los mecanismos existentes de diafonía entre rutas biológicas en los organismos. Más específicamente, se buscó desarrollar metodologías computacionales que asistan en la reconstrucción (o descubrimiento) de la estructura relacional presente en las redes de rutas biológicas.

Una ruta biológica representa una secuencia de reacciones o interacciones entre un grupo de genes expresados que participan en un proceso biológico. Los estudios que se centran sólo a nivel de genes difícilmente pueden explicar muchos fenómenos celulares complejos, por lo que un enfoque para construir o descubrir las diafonías entre las rutas biológicas se ha convertido en una estrategia clave en las últimas décadas. Esta *diafonía* se refiere a la situación en la que uno o más componentes de una ruta afectan a otra, esto genera una coordinación entre los procesos biológicos descritos por cada ruta.

En esta tesis se presentó como principal aporte un método bautizado como PET (*Crosstalk **P**athway Inference by using Gene **E**xpression Data Biclustering and **T**opological Information*) que busca inferir redes de rutas biológicas basándose tanto en la topología como en los datos de expresión de genes resultado de un experimento de control vs. afectados mediante el uso clave de biclustering. Esto permite detectar conexiones entre rutas que no comparten genes en común, es decir relaciones no tan obvias que no son encontradas por

otros métodos. De este modo se obtuvo una metodología que facilita la interpretabilidad de la relevancia biológica de las interacciones encontradas.

El rendimiento de esta metodología se contrastó con dos algoritmos(Dutta, Wallqvist, & Reifman, 2012)(Ponzoni, y otros, 2014) con similar objetivo, utilizando datos correspondientes a la enfermedad de Alzheimer (AD). Las diafonías inferidas por PET capturaron la mayoría de las interacciones detectadas por estos otros métodos, pero además PET detectó varias nuevas relaciones no encontradas por ellos. Estos resultados fueron analizados con más detalle en el capítulo 7. Además se realizó un análisis ontológico mostrando que los biclusters de consenso detectados por la herramienta tienen una relación bien establecida en términos de anotaciones funcionales.

A partir de estos resultados, es posible concluir que las ideas exploradas en esta tesis constituyen una iniciativa valiosa para el diseño de un enfoque integrador para la detección de diafonía entre rutas biológicas. En el futuro, esperamos impulsar nuevos experimentos e incorporar estrategias adicionales para mejorar la inferencia de señales de diafonía. En particular, las nuevas métricas para la integración y la conciliación de las diferentes fuentes de evidencia diafonía están en desarrollo.

8.1 Contribuciones

Biclustering con tecnología de microarray

La tecnología de microarrays constituye uno de los métodos más utilizados que contribuyen a generar una gran cantidad de datos biológicos, como se ha visto en este trabajo de tesis es limitada la capacidad para extraer información biológica significativa de estos datos. En este sentido, el reconocimiento de los grupos de genes con valores de expresión coherentes representa un paso clave en el análisis de datos de expresión génica. Como se ha tratado con más detalle en el capítulo 4, los algoritmos de agrupación tradicionales intentan encontrar patrones de expresión en submatrices que se extienden sobre todo el conjunto de muestras (clustering), lo que es demasiado restrictivo suponer que todos los genes se comportan de manera similar en todas las condiciones. Teniendo esto en cuenta, se llevan a cabo los enfoques de biclustering, que agrupan en ambas dimensiones simultáneamente: genes y muestras (Madeira & Oliveira, 2004)(Madeira & Oliveira, 2009). Se han propuesto varios algoritmos de biclustering en la literatura(Cheng & Church, 2000)(DiMaggio, McAllister, Floudas, Feng, Rabinowitz, & Rabitz, 2008)(Mitra & Banka, 2006)(Bleuler, Prelic, & Zitzler, 2004)(Gallo, Carballido, & Ponzoni, 2009) variando de enfoques codiciosos simples a algoritmos evolutivos estocásticos complejos. Sin embargo el software que los implementa es difícil de usar o no se encuentra accesible. Por esta razón se presenta una herramienta con el fin de lograr el uso de la técnica de BiHEA (Gallo, Carballido, & Ponzoni, 2009) denominada BAT (por sus siglas en inglés, ***B****iclustering* ***A****nalysis* ***T****oolbox*) (Gallo, Dussaut, Carballido, & Ponzoni, 2010). La herramienta aporta una

ayuda al bioinformático en temas de pre-procesamiento de datos, ejecución del algoritmo BiHEA, cambio de sus parámetros, visualización en varios niveles y post-procesamiento en conjunto con la posibilidad de guardar el proyecto o los diferentes elementos que se han utilizado y procesado.

Minería de texto

Desde la perspectiva de la minería de texto, las rutas biológicas son conocimientos bien establecidos acerca de los procesos biológicos, a partir de interpretaciones de un gran número de hechos dispersos a través de la literatura (Oda, y otros, 2008)(Buyko, Linde, Priebe, & Hahn, 2011), donde la publicación rápida de nuevos documentos hacen que estar al día sea un grave problema (por ejemplo, la base de datos de PubMed contiene información de más de 25 millones de artículos y continúa creciendo a un ritmo elevado cada semana). Por lo tanto, los métodos de minería de texto, que ayudan en la construcción y mantenimiento de los conocimientos sobre las rutas, se han convertido en herramientas pertinentes para los biólogos para gestionar esta cantidad cada vez mayor de la literatura biológica. Sin embargo, el descubrimiento de las rutas biológicas en la minería de texto no es una tarea fácil. El conocimiento significativo es muy difícil de extraer automáticamente. Otro tema crucial en la minería de texto aplicados a la Bioinformática es lograr una validación sólida de los métodos debido a la falta de grandes conjuntos de pruebas, validación objetiva o "normas de oro" (Maguitman, Rechtsteiner, Verspoor, Strauss, & Rocha, 2006). Estos problemas tienen como principal consecuencia que muchas rutas inferidas no representan explicaciones coherentes de los hechos denunciados (Li, Liakata, & Rebholz-Schuhmann, 2013), y para transformar los resultados de las redes construidas automáticamente en rutas biológicas parece requerir importantes esfuerzos

humanos adicionales. Por esa razón, la integración de algoritmos de minería de la literatura con las estrategias de validación robustas para la extracción de conocimiento sobre rutas biológicas es un interesante campo de investigación abierto.

En esta tesis se presentó una enfoque de minería de texto para asistir en la construcción de una red de rutas biológicas denominada PANTex(Dussaut, Cravero, Ponzoni, Maguitman, & Cecchini, 2014). Previamente se desarrolló una metodología integrativa para la clasificación de genes en rutas biológicas siguiendo el mismo enfoque(Dussaut, Ponzoni, Cecchini, & Maguitman, 2012). Este método aprovecha las ventajas de un algoritmo de predicción introducido en Maguitman et al. (Maguitman, Rechtsteiner, Verspoor, Strauss, & Rocha, 2006). El algoritmo propuesto utiliza un corpus llamado BioCreative(Krallinger, Leitner, & Valencia, 2007) y una herramienta denominada ABNER (Settles, 2005) para identificar entidades biológicas mencionadas en las publicaciones. Se crea una matriz de frecuencia de genes, que es utilizada para calcular un peso de co-ocurrencia para cada par de genes. La predicción se hace entonces basada en un sistema de votación adaptado del algoritmo de Maguitman. Para validar los resultados se utilizó la información de la base de datos de KEGG (Kanehisa, Goto, Sato, Kawashima, Furumichi, & Tanabe, 2014)(Kanehisa & Goto, 2000) midiendo precisión y cobertura. Partiendo de esta base, se desarrolló PANTex (Dussaut, Cravero, Ponzoni, Maguitman, & Cecchini, 2014), como punto de partida de este enfoque. Este método utiliza la base de datos KEGG con el fin de reunir una lista de rutas biológicas para cada organismo, en la etapa inicial se consideraron como organismos válidos solamente humanos y levadura. Utilizando esta lista buscamos publicaciones PubMed a través de su programación *Entrez Utilities*(NCBI Resource Coordinators, 2013) y detectamos la co-ocurrencia de las rutas biológicas en la misma publicación. Los datos resultantes se almacenan en una matriz de intersección. También mantenemos un registro del número de publicaciones que contienen cada nombre de ruta para fines de normalización. Con el fin de validar el método propuesto contrastamos la matriz

normalizada resultante con los datos reportados en Alexeyenko y Sonnhammer (Alexeyenko & Sonnhammer, 2009). Este método es tratado con más detalle en el capítulo 6.

Diafonía de rutas biológicas

La principal motivación para identificar diafonías entre las rutas biológicas, es la comprensión de los fenómenos biológicos que se producen a nivel celular, ya que la coordinación o diafonía entre los procesos biológicos resulta muy compleja en un nivel inferior (genes que subyacen en las diferentes rutas). Se han propuesto varios enfoques abordados en el capítulo 6 de esta tesis. En el capítulo 7 se detalla el aporte central de la tesis denominado PET (por sus siglas en inglés *Crosstalk **P**athway Inference by using Gene **E**xpression Data Biclustering and **T**opological Information*). Este último enfoque utiliza la combinación del análisis de enriquecimiento basándose en la información topológica obtenida de KEGG (Kanehisa, Goto, Sato, Kawashima, Furumichi, & Tanabe, 2014)(Kanehisa & Goto, 2000)y biclustering de datos de expresión genética. En pocas palabras, dado un conjunto de datos de microarrays y un par de rutas en estudio, el análisis de enriquecimiento identifica qué partes de las rutas son co-activadas por los genes con expresión diferencial en los conjuntos de datos, mientras que el análisis biclustering captura la sincronización entre los perfiles de expresión de estos genes. El rendimiento de este método se contrastó con otros dos arrojando resultados positivos detallados en el capítulo 7.

8.2 Trabajo futuro

Como extensión natural a las investigaciones realizadas en esta tesis, se desean continuar mejorando los métodos desarrollados. Con respecto al campo de la minería de datos se está trabajando conjuntamente con la Dra. Ana Maguitman y la Dra. Rocío Cecchini en el desarrollo de una técnica que busca generar una red derivada de la co-citación de rutas biológicas. En el ámbito de encontrar diafonías entre rutas biológicas, esperamos impulsar nuevos experimentos e incorporar estrategias adicionales para mejorar la captura de señales de diafonía. En particular, las nuevas métricas para la integración y la conciliación de las diferentes fuentes de evidencia diafonía están en desarrollo. También se prevé explorar la integración en el proceso de inferencia de conocimiento biológico externo correspondiente a bases de datos como KEGG y Gene Ontology, e incluso integrar otras fuentes de datos biológicos que no necesariamente representan niveles de expresión de genes, a fin de refinar aún más los conocimientos extraídos.

8.3 Publicaciones

1. Gallo C.A., **Dussaut J.S.**, Carballido J.A., Ponzoni I. "BAT: A new Biclustering Analysis Toolbox", In: Ferreira, C.E.; Miyano, S.; Stadler, P.F. (Eds.): Advances in Bioinformatics and Computational Biology, 5th Brazilian Symposium on Bioinformatics, BSB 2010, Buzios, Rio de Janeiro, Brazil, August 30-September 3, 2010, Proceedings. Lecture Notes in Computer Science Vol. 6268, pp. 67-71. Springer-Verlag Berlin Heidelberg, (2010). ISSN: 0302-9743. Indexada en DBLP, Scopus e ISI Proceedings.

2. Gallo C.A., **Dussaut J.S.**, Carballido J.A., Ponzoni, I. “A Microarray Biclustering Analysis Tool based on the BiHEA Algorithm”, CAB2C 2010 (I Congreso Argentino de Bioinformática y Biología Computacional), 12-14 de mayo de 2010, Quilmes, Argentina.

3. **Dussaut J.S.** “BAT: A new Biclustering Analysis Toolbox”, Monografía de Proyecto Final de Carrera de Ingeniería en Sistemas de Computación, Depto. Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur. Bahía Blanca, Argentina. Agosto de 2010.

4. **Dussaut J.S.**, Ponzoni I., Cecchini R.L., Maguitman A. “An integrative methodology for the classification of genes into pathways using a novel text mining approach”, ISCB-LA 2012, 17-21 de marzo de 2012, Santiago, Chile.

5. **Dussaut J.S.**, Ponzoni I., Cecchini R.L., Maguitman A. “PaNTex: A novel methodology to assemble Pathway Networks using Text Mining”, Bariloche V CAB2C, 21-24 de septiembre de 2014, Bariloche, Argentina.

6. Ponzoni I., Nueda M.J., Tarazona S., Götz S., Montaner D., **Dussaut J.S.**, Dopazo J., Conesa A. “Pathway network inference from gene expression data”, BMC Systems Biology 2014, 8(Suppl 2):S7.

7. Carballido J.A., Gallo C.A., **Dussaut J.S.**, Ponzoni I. “On Evolutionary Algorithms for Biclustering of Gene Expression Data”, Current Bioinformatics 2015, 10(3): 259-267.

8. **Dussaut J.S.**, Gallo C.A., Cecchini R.L., Carballido J.A., Ponzoni I, “Crosstalk Pathway Inference using Topological Information and Biclustering of Gene Expression Data”, under review.

Referencias

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM-SIGMOD International Conference on Management of Data* (págs. 207-216). Washington, DC, USA: ACM Press.

Aguilar-Ruiz, J. S. (2005). Shifting and scaling patterns from gene expression data. *Bioinformatics, 21* (20), 3840-3845.

Alberts, B., Johnson, A., Lewis, J., Raff, M., & Roberts, K. (2004). *Biología Molecular de la Célula.* Barcelona: Omega.

Alexeyenko, A., & Sonnhammer, E. (2009). Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome Research, 19*, 1107-1116.

Amaratunga, D., & Cabrera, J. (2009). Exploration and Analysis of DNA Microarray and Protein Array Data. *Wiley Series in Probability and Statistics* .

Aréchiga, H. (1996). *Los fenómenos fundamentales de la vida.*Siglo XXI.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., y otros. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics, 25* (1), 25-29.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463). New York: ACM press.

Bartus , K., Pigott , B., & Garthwaite , J. (2013). Cellular targets of nitric oxide in the hippocampus. *PLoS One, 8*, 57292.

Ben-Dor, A., Chor, B., Karp, R., & Yakhini, Z. (2002). Discovering local structure in gene expression data: The order–preserving submatrix problem. *Proceedings of the 6th International Conference on Computacional Biology (RECOMB'02)* , 49–57.

Bleuler, S., Prelic, A., & Zitzler, E. (2004). An EA framework for biclustering of gene expression data. *Proceeding of Congress on Evolutionary Computation* , 166-173.

Bremermann, H. J. (1962). Optimization through. *Self-Organizing systems* , 93–106.

Brini , M., & Carafoli , E. (2009). Calcium Pumps in Health and Disease. *Physicological Reviews, 89*, 1341–1378.

Brzyska , M., & Elbaum , D. (2003). Dysregulation of calcium in Alzheimer's disease. *Acta Neurobiologiae Experimentalis, 63*, 171-183.

Buyko, E., Linde, J., Priebe, S., & Hahn, U. (2011). Towards automatic pathway generation from biological full-text. *Lecture Notes in Computer Science, 7014*, 67-79.

Carballido, J. A., Gallo, C. A., Dussaut, J. S., & Ponzoni, I. (2015). On Evolutionary Algorithms for Biclustering of Gene Expression Data. *Current Bioinformatics, 10* (3), 259-267.

Chang, T. W. (1983). Binding of cells to matrixes of distinct antibodies coated on solid surface. *Journal of immunological methods, 65* (1), 217-223.

Cheng, Y., & Church, G. M. (2000). Biclustering of expression data. *Ismb, 8*, 93-103.

Chiocco , M. J., Zhu , X., Walther , D., Pletnikova , O., Troncoso , J. C., Uhl , G. R., y otros. (2010). Fine mapping of calcineurin (PPP3CA) gene reveals novel alternative splicing patterns, association of 5'UTR trinucleotide repeat with addiction vulnerability, and differential isoform expression in Alzheimer's disease. *Substance Use Misuse, 45* (11), 1809-26.

Coello, C. C., Van Veldhuizen, D. A., & Lamont, G. B. (2002). *Evolutionary algorithms for solving multi-objective problems.* New York: Kluwer Academic.

Cruz , N. F., Ball , K. K., & Dienel , G. A. (2010). Astrocytic gap junctional communication is reduced in amyloid-β-treated cultured astrocytes, but not in Alzheimer's disease transgenic mice. *ASN Neuro, 2* (4), 201–213.

Das, D. B., & Patvardhan, C. (1998). New multi-objective stochastic search technique for economic load dispatch. *Generation, Transmission and Distribution, IEE Proceedings, 145* (6), 747-752.

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. A. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *Evolutionary Computation, IEEE Transactions on, 6* (2), 182-197.

DiMaggio, P., McAllister, S., Floudas, C., Feng, X. J., Rabinowitz, J., & Rabitz, H. (2008). Biclustering via optimal re-ordering of data matrices in systems biology: rigorous methods and comparative studies. *BMC Bioinformatics, 9* (1), 458.

Divina, F., & Aguilar-Ruiz, J. S. (2006). Biclustering of Expression Data with Evolutionary Computation. *IEEE Trans Knowl Data Eng, 18* (5), 590-602.

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., y otros. (2007). A systems biology approach for pathway level analysis. *Genome research, 17* (10), 1537-1545.

Dussaut, J. S., Cravero, F., Ponzoni, I., Maguitman, A. G., & Cecchini, R. L. (2014). PaNTex: A novel methodology to assemble Pathway Networks using Text Mining. *VCAB2C*, (págs. 38-41). Bariloche, Argentina.

Dussaut, J. S., Ponzoni, I., Cecchini, R. L., & Maguitman, A. G. (2012). An integrative methodology for the classification of genes into pathways using a novel text mining approach. *ISCB-LA.* Santiago, Chile.

Dutta, B., Wallqvist, A., & Reifman, J. (2012). PathNet: a tool for pathway analysis using topological information. *Source Code for Biology and Medicine, 7* (10).

Fogel, L. J., Owens, A. J., & Walsh, M. J. (1966). Intelligent decision making through a simulation of evolution. *Behavioral science, 11* (4), 253-272.

Francesconi, M., Remondini, D., Neretti, N., Sedivy, J. M., Cooper, L. N., Verondini, E., y otros. (2008). Reconstructing networks of pathways via significance analysis of their intersections. *BMC bioinformatics, 9*.

Fraser, A. S. (1958). Simulation of Genetic Systems by Automatic Digital Computers IV. Selection Between Alleles at a Sex-Linked Locus. *Australian Journal of Biological Sciences, 11* (4), 613-626.

Gallo, C. A., Carballido, J. A., & Ponzoni, I. (2009). BiHEA: A Hybrid Evolutionary Approach for Microarray Biclustering. *Lecture Notes in Computer Science, 5676*, 36-47.

Gallo, C. A., Dussaut, J. S., Carballido, J. A., & Ponzoni, I. (2010). BAT: a new biclustering analysis toolbox. En *Advances in Bioinformatics and Computational Biology* (págs. 67-70). Berlin Heidelberg: Springer.

Gasch, A. P., & Eisen, M. B. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology, 3* (11).

Getz, G., Levine, E., & Domany, E. (2000). Coupled two-way clustering analysis. *Proceedings of the NaturalAcademy of Sciences USA* , 12079–12084.

Goeman, J. J., Van De Geer, S. A., De Kort, F., & Van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics, 20* (1), 93-99.

Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the American, 67* (337), 123–129.

Hermes , M., Eichhoff , G., & Garaschuk , O. (2010). Intracellular calcium signalling in Alzheimer's disease. *Journal of Cellular and Molecular Medicine, 14* (1-2), 30-41.

Holland, J. H. (1975). Adaptation in Natural and Artificial Systems. *The University of Michigan Press* .

Holland, J. H. (1962). Outline for a logical theory of adaptive systems. *JACM, 9* (3), 297–314.

Hsu, C. L., & Yang , U. C. (2012). Discovering pathway cross-talks based on functional relations between pathways. *BMC Genomics, 13* (7).

Huang , D. W., Sherman, B. T., & Lempicki , R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research, 37* (1), 1-13.

Huang , D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protocol, 4* (1), 44-57.

Irizarry, R. A., Wang, C., Zhou, Y., & Speed, T. P. (2009). Gene Set Enrichment Analysis Made Simple. *Statistical Methods in Medical Research, 18* (6), 565-75.

Jin, Y., Olhofer, M., & Sendhoff, B. (2001). Dynamic weighted aggregation for evolutionary multi-objective optimization: Why does it work and how?

Joung, J. G., Kim, S. J., Shin, S. Y., & Zhang, B. T. (2012). A probabilistic coevolutionary biclustering algorithm for discovering coherent patterns in gene expression dataset. *BMC Bioinformatics, 13* (17), 12.

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

Kanehisa , M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research, 42*, 199–205.

Karp, G. (2009). *Cell and Molecular Biology: Concepts and Experiments.* John Wiley & Sons.

Kluger, Y., Basri, R., Chang, J., & Gerstein, M. (2003). Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research, 13*, 703–716.

Krallinger, M., Leitner, F., & Valencia, A. (2007). Assessment of the second BioCreative PPI task: automatic extraction of protein-protein interactions. *Second bioCreative Challenge Evaluation Workshop.*

LaFerla , F. M. (2002). Calcium dyshomeostasis and intracellular signalling in Alzheimer's disease. *Nature Reviews Neuroscience, 3* (11), 862-72.

Li, C., Liakata, M., & Rebholz-Schuhmann, D. (2013). Biological network extraction from scientific literature: state of the art and challenges. *Briefings in bioinformatics* .

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association, 62* (318), 399-402.

Liu, Z. P., Wang , Y., Zhang , X. S., & Chen , L. (2010). Identifying dysfunctional crosstalk of pathways in various regions of Alzheimer's disease brains. *BMC Systems Biology, 4*.

Lu, L. J., Sboner, A., Huang, Y. J., Lu, H. X., Gianoulis, T. A., Yip, K. Y., y otros. (2007). Comparing classical pathways and modern networks: towards the development of an edge ontology. *Trends in Biochemical Sciences, 32*, 320-331.

Madeira, S., & Oliveira, A. (2009). A polynomial time biclustering algorithm for finding approximate expression patterns in gene expression time series. *Algorithms for Molecular Biology, 4* (1), 8.

Madeira, S., & Oliveira, A. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE-ACM Transactions on Computational Biology and Bioinformatics, 1*, 24–45.

Maguitman, A. G., Rechtsteiner, A., Verspoor, K., Strauss, C., & Rocha, L. (2006). Large-Scale Testing of Bibliome Informatics. *Pacific Symposium on Biocomputing*, (págs. 76-87).

Marambaud , P., Dreses-Werringloer, U., & Vingtdeux , V. (2009). Calcium signaling in neurodegeneration. *Molecular Neurodegeneration, 4*, 20.

Mattson , M. P., & Chana , S. L. (2003). Neuronal and glial calcium signaling in Alzheimer's disease. *Cell Calcium, 34*, 385–397.

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences, 74* (2), 560-564.

Mcdonald, D. M., Chen, H., Su, H., & Marshall, B. B. (2004). Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser. *Bioinformatics, 20* (18), 3370-3378.

Mei , X., Ezan , P., Giaume , C., & Koulakoff , A. (2010). Astroglial connexin immunoreactivity is specifically altered at β-amyloid plaques in beta-amyloid precursor protein/presenilin1 mice. *Neuroscience, 171* (1), 92–105.

Mitra, S., & Banka, H. (2006). Multi-objective evolutionary biclustering of gene expression data. *Pattern Recognition, 39*, 2464-2477.

Monfort , P., Muñoz , M. D., Kosenko , E., Llansola , M., Sánchez-Pérez , A., Cauli , O., y otros. (2004). Sequential activation of soluble guanylate cyclase, protein kinase G and cGMPdegrading phosphodiesterase is necessary for proper induction of long-term potentiation in CA1 of hippocampus. Alterations in hyperammonemia. *Neurochemistry Int, 45*, 895–901.

Mukhopadhyay, A., Maulik, U., & Bandyopadhyay, S. (2008). Evolving coherent and non-trivial biclusters from gene expression data: An evolutionary approach. *Proc IEEE Region 10 Conf 2008.*

NCBI Resource Coordinators. (2013). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research, 41*, 8-20.

Oda, K., Kim, J. D., Ohta, T., Okanohara, D., Matsuzaki, T., Tateisi, Y., y otros. (2008). New challenges for text mining: mapping. *BMC Bioinformatics, 9* (3), 5.

Orsetti , M., Di Brisco, F., Canonico , P. L., Genazzani , A. A., & Ghi , P. (2008). Gene regulation in the frontal cortex of rats exposed to the chronic mild stress paradigm, an animal model of human depression. *European Journal of Neuroscience, 27*, 2156–2164.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The pagerank citation ranking: Bringing order to the Web.*Palo Alto, CA: Stanford University.

Patergnani , S., Marchi , S., Rimessi , A., Bonora , M., Giorgi , C., Mehta , K. D., y otros. (2013). PRKCB/protein kinase C, beta and the mitochondrial axis as key regulators of autophagy. *Autophagy, 9* (9), 1367-1385.

Pontes, B., Divina, F., Giráldez, R., & Aguilar-Ruiz, J. S. (2013). Configurable pattern-based evolutionary biclustering of gene expression data. *Algorithms for Molecular Biology, 8* (4).

Pontes, B., Divina, F., Giráldez, R., & Aguilar-Ruiz, J. S. (2010). Measuring the Quality of Shifting and Scaling Patterns in Biclusters. *Pattern Recognition in Bioinformatics, Lecture Notes in Computer Science, 6282*, 242-252.

Pontes, B., Divina, F., Giráldez, R., & Aguilar-Ruiz, J. S. (2007). Virtual Error: A New Measure for Evolutionary Biclustering. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, Lecture Notes in Computer Science 2010, 4447*, 217-226.

Ponzoni, I., Nueda, M. J., Tarazona, S., Götz, S., Montaner, D., Dussaut, J. S., y otros. (2014). Pathway network inference from gene expression data. *BMC Systems Biology, 8*.

Rechenberg, I. (1971). Evolutionsstrategie - Optimierung technischer Systeme nach Prinzipien der biologischen Evolution. *PhD Thesis* .

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences, 74* (12), 5463-5467.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., y otros. (2009). PID: the Pathway Interaction Database. *Nucleic Acids Research, 37*, 674–679.

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science, 270* (5235), 467-470.

Schwefel, H. P. (1974). Numerische Optimierung von Computer-Modellen. *PhD Thesis* .

Settles, B. (2005). ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics, 21*, 3191-3192.

Setubal, J. C., & Meidanis, J. (1997). *Introduction to computational molecular biology.* PWS Publishing.

Shen , F., Li , Y. J., Shou , X. J., & Cui , C. L. (2012). Role of the NO/sGC/PKG signaling pathway of hippocampal CA1 in morphine-induced reward memory. *Neurobiology of Learning and Memory, 98*, 130–138.

Standafer, E., & Wahlgren, W. (2002). *Modern Biology.* Holt, Rinehart and Winston.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., y otros. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting

genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America, 102* (43), 15545-15550.

Tanay, A., Sharan, R., & Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics, 18*, 136–144.

Tang, C., Zhang, A., Zhang, L., & Ramanathan, M. (2001). Interrelated two-way clustering: An unsupervised approach for gene expression data analysis. *Proceeding of BIBE2001: 2nd IEEE International Symposium on Bioinformatics and Bioengineering* , 41–48.

Thomas , R., Gohlke, J. M., Stopper, G. F., Parham, F. M., & Portier, C. J. (2009). Choosing the right path: enhancement of biologically relevant sets of genes or proteins using pathway structure. *Genome Biology* .

Uddin , R. K., & Singh , S. M. (2013). Hippocampal Gene Expression Meta-Analysis Identifies Aging and Age-Associated Spatial Learning Impairment (ASLI) Genes and Pathways. *PLoS ONE, 8* (7), 69768.

Weiss, S. M., & Indurkhya, N. (1998). *Predictive Data Mining.* Morgan Kaufmann.

Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering, 12* (3), 372-390.

Zhao, H., Liew, A. W., Wang, D. Z., & Yan, H. (2012). Biclustering Analysis for Pattern Discovery: Current Techniques, Comparative Studies and Applications. *Current Bioinformatics, 7* (1), 43-55.

Zündorf , G., & Reiser , G. (2011). Calcium Dysregulation and Homeostasis of Neural Calcium in the Molecular Mechanisms of Neurodegenerative Diseases Provide Multiple Targets for Neuroprotection. *Antioxidants and Redox Signaling, 14* (7), 1275-88.