12

Análisis de datos de expresión génica: de la información al conocimiento en la Biología actual

Isabel Nepomuceno Chamorro
Juan A. Nepomuceno Chamorro

12.1. Justificación

Una posible justificación para enmarcar este trabajo en el campo de la LLI (Lógica, Lenguaje e Información) podría ser argumentar que la *Bioinformática* es la disciplina que, mediante el uso de técnicas computacionales (léase tanto a nivel teórico como tecnológico), se encarga de procesar la *información* biológica obtenida para generar el conocimiento que nos permita comprender la *lógica* subyacente en los sistemas biológicos. Todo ello enmarcado en el contexto del Proyecto Genoma Humano y aquella metáfora, tan recurrente, que se solía emplear para explicar que el estudio del genoma humano consistía en la decodificación del *lenguaje* en el que se escribía el libro de la Vida.

Sin embargo, aunque puede llegar a resultar convincente, la verdadera razón para enmarcar este trabajo en un volumen sobre "temas lógicos" es de otro tipo. Los autores de este trabajo[1] frecuentemente son preguntados por sus resultados, la motivación de sus investigaciones y su trabajo en general por el homenajeado. Ambos autores se dedican al estudio y análisis de datos de expresión génica desde una perspectiva de la Minería de Datos, cada cual en un campo diferente. A grandes rasgos, sus trabajos pueden enmarcarse dentro de la Informática Aplicada al estudio de problemas de tipo biológico. Este artículo, de carácter divulgativo, presenta la motivación general y el marco en el que se encuadran sus trabajos de investigación.

12.2. La Biología a comienzos del siglo XXI

Según la visión clásica o reduccionista de la Vida, toda la información necesaria para la génesis, desarrollo y funcionamiento de un organismo se encuentra almacenada en las

[1]El lector puede observar los apellidos de los autores y elaborar su propia hipótesis.

moléculas de ADN. Estas moléculas constituyen el código genético o material hereditario. Bajo este enfoque, se puede afirmar que estas moléculas encierran las instrucciones de la vida que son codificadas en un lenguaje de cuatro letras: A, G, C y U. Se denomina genoma al material genético contenido en las células de un organismo. En las células de mayor complejidad estructural o eucariotas, el término genoma se refiere al ADN almacenado en el núcleo de la misma. Tras el Proyecto Genoma Humano surge un cambio de paradigma, donde la Vida se considera como un Sistema Complejo donde las partes, ya sean a nivel genético, proteico o de metabolitos, están continuamente interactuando entre sí y el medio, para dar lugar a las funciones biológicas.

La *Bioinformática* y *Biología Computacional* son disciplinas que surgen, una vez secuenciado el genoma humano, como disciplinas puentes entre la información generada, el conocimiento biológico y la práctica médica. Estas disciplinas tratan de proporcionar modelos de diagnósticos tempranos, de tratamientos, etc, personalizados a cada paciente.

12.2.1. Visión causal de la Vida: el reduccionismo

La clasificación de los organismos vivos se realiza en función: del número de células que posea o según la complejidad estructural de las mismas. La clasificación, según el número de células, es en organismos unicelulares (bacterias y protozoos) o pluricelulares. La clasificación, según la complejidad estructural de las células, es en organismos procariotas o eucariotas. Estos organismos difieren en si el ADN se encuentra almacenado en un compartimiento nuclear diferenciado (eucariotas) o no (procariotas). Las células contienen el material genético o hereditario o ADN distribuido en cromosomas y genes. Éste material es el responsable de la síntesis de proteínas. Cualquier organismo pluricelular, como por ejemplo el ser humano, está constituido a partir de más de 100,000 proteínas. De este modo, podemos considerar la célula como una gran fábrica que se encuentra en constante producción y en la que se produce sucesivamente una secuencia gigantesca de relaciones o interacciones entre moléculas. Esta secuencia de interacciones es lo que se denominan rutas bioquímicas o *pathways*. Las rutas bioquímicas son de diferente índole, como por ejemplo las rutas relacionadas con el desarrollo o el ciclo celular. Pero en todos los casos las rutas reguladoras tienen como resultado la síntesis de proteínas. Si observamos las células como colonias de individuos, podemos decir que éstas son causa de los tejidos ya que se agrupan formándolos. Las células se especializan y así forman diferentes tejidos como son el óseo, el muscular, etc. Estos tejidos constituyen los órganos que son elementos fundamentes como el corazón o el riñón. Y éstos junto con el sistema circulatorio, el sistema inmune y otros constituyen el organismo, es decir, son causa del ser vivo.

Bajo este enfoque reduccionista, la Vida puede considerarse como una cadena causal en la que el fenotipo o función se crea a partir del genotipo o información genética, como se acaba de describir. Esta tendencia se ha desarrollado teniendo como eje de todo el proceso los genes y fue acuñada por el filósofo André Pichot como *ADNmanía*. La ADNmanía es el contexto bajo el que se desarrollo el Proyecto Genoma Humano.

12.2.2. El Proyecto Genoma Humano

A mediados de la década de los 80, los avances en Genética y Biología Molecular proporcionaban las herramientas indispensables para conocer en su totalidad el genoma humano, así como el genoma de otras especies consideradas organismos modelos. Esta

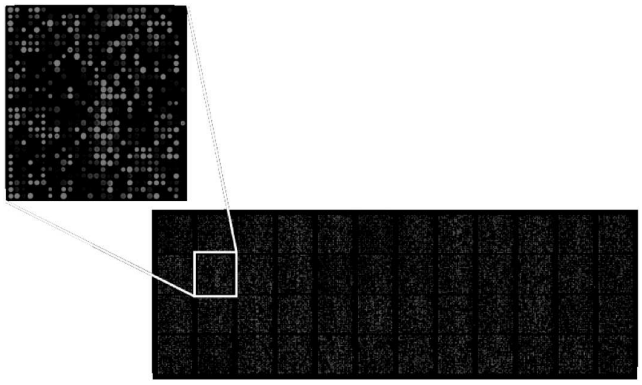

Figura 12.1: Resultado del escaneado de microarray

iniciativa surgió con el apoyo económico del Congreso de los Estados Unidos y fue administrada a través del Instituto Nacional de Salud[2] y el Departamento de Energía de este país. Más tarde esta empresa se convertiría en una iniciativa a nivel mundial cuya estructura dio origen a la Organización Genoma Humano (Human Genome Organization, HUGO). El objetivo del Proyecto Genoma Humano se desarrollaría a través de dos vías: la secuenciación y el mapeo genético. Mediante la secuenciación se trata de averiguar la posición de todos los nucleótidos o letras (A,T,U y C) del lenguaje que constituye el genoma. Mientras que con el mapeo genético se trata de cartografiar, es decir, de localizar los genes en cada uno de los 23 pares de cromosomas o componentes en los que se compacta el ADN y por tanto genes del ser humano.

Tras el Proyecto Genoma Humano, la aparición y el uso de técnicas computacionales son de vital relevancia para que la información extraída en laboratorio se convierta en conocimiento biológico que repercuta en el entorno clínico. La Bioinformática y la Biología Computacional juega hoy en día un rol relevante en la investigación conocida como *Translational Biomedical Research*, ya que son herramientas puente entre el la información en biología y la práctica médica.

12.2.3. Tecnología de Microarray y Análisis de Datos de Expresión Génica

La *técnica de microarray* también se conoce con los nombres de *biochip*, *DNA chip* o *array de genes*, pero el término más utilizado es el de microarray. Esta tecnología se fundamenta en la hibridación molecular y obtiene como resultado una matriz de puntos a modo de puzzle en el que la lectura de presencia o ausencia de un color se identifica con el gen expresado o no (ver Figura 12.1). De este modo se puede obtener la huella genética o cuadro genómico de la muestra bajo estudio. Esta tecnología ha revolucionado la investigación en Biología por su capacidad de monitorizar el nivel de expresión de miles de genes en simultáneo [1].

El análisis de un experimento de microarray requiere de una serie de etapas a realizar que abarcan desde establecer la hipótesis inicial, en nuestro caso sería acerca de una cuestión biológica o clínica, hasta la validación en laboratorio del modelo generado

[2]Por suerte dicho instituto no tenía sede en nuestra comunidad pues, en ese caso, estaría aún en trámites para su creación.

computacionalmente a partir de los datos, véase la figura 12.2. Tras fijar la hipótesis de partida las etapas a seguir son:

- **Diseño del experimento**, donde los investigadores deciden qué tipo de tecnología de microarray va a ser utilizada (cADN o microarray de oligonucleótidos), cómo se van a hibridizar los chips, etc.

- La segunda etapa se denomina en la literatura ***image scanning process*** que consiste en extraer o quantificar la información que contiene las sondas de microarray. Tras esta etapa se obtiene los datos en crudo. Esta etapa se suele incluir también como parte de la siguiente etapa.

- A continuación se realiza un **preprocesado** de los datos, es la etapa que se denomina *Low Level Analysis*. En esta etapa se tiene en cuenta las diferentes condiciones (diferentes pacientes, días, laboratorios, intensidad del fluorescente, etc.) del microarray que pueden dar lugar a diferentes rangos de expresión. En esta etapa se suelen aplicar técnicas de normalización y escalado a los datos debido a que las diferentes condiciones del microarray no suelen tener una escala comparable.

- La siguiente etapa consiste en realizar tareas de **Minería de Datos**, es la etapa denominada *High Level Analysis*. En esta etapa es donde se encuadra la realización de nuestra investigación. El objetivo será crear modelos que describan el comportamiento a nivel genético de un conjunto de muestras.

- Finalmente, la última etapa se conoce como ***Determination of biological significance*** y ***Biological Verification***. Es aquí donde se analizan los resultados proporcionados por la etapa anterior para así estimar la relevancia biológica de dichos resultados. El estudio de significancia biológica puede encuadrarse también como parte final de la etapa de análisis de alto nivel. El análisis de significancia se realiza a partir de una lista de genes resultado de la etapa anterior y utilizando repositorios públicos. En cuanto verificación hacemos referencia a la verificación en laboratorio de las conclusiones obtenidas a partir del análisis de alto nivel.

12.2.4. Análisis de Alto Nivel

El análisis de alto nivel consiste en realizar tareas de Minería de Datos y Aprendizaje Automático con el objetivo de extraer conocimiento a partir de la información que nos proporciona un experimento de microarray: relaciones entre genes, patrones de comportamineto o, en general, modelos de conocimiento. Las técnicas de minería están dirigidas, en el análisis de microarray, a la identificación de genes con un comportamiento similar dentro de un conjunto de condiciones. Básicamente es un proceso de selección de genes que incluye tareas de filtrado y creación de modelos de comportamiento. El análisis de alto nivel enmarca tareas como las de:

- *Clustering*: agrupamiento de genes con un mismo comportamiento de manera global.

- *Biclustering*: agrupamientos de genes con un mismo comportamiento de manera local (tema de investigación de Juan Antonio).

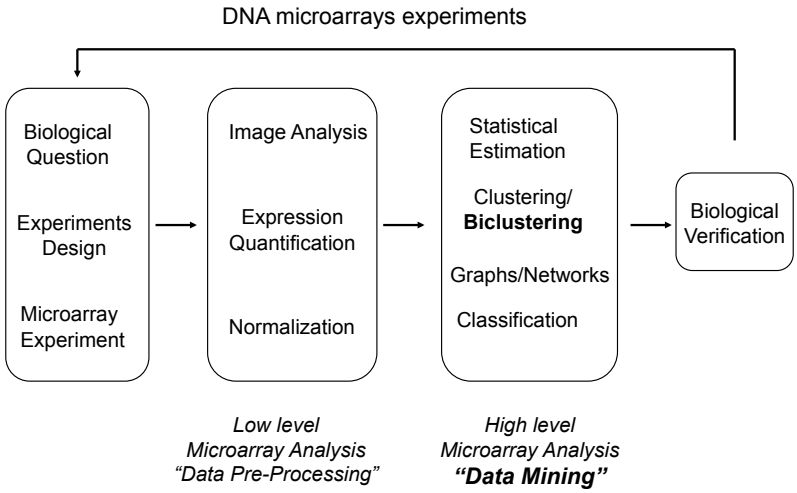

Figura 12.2: Análisis de Microarray

- Redes de asociación de genes: grafos de dependencias entre genes (tema de investigación de Isabel).
- Reglas de clasificación, etc.

El objetivo de nuestra investigación se enmarca dentro de esta etapa y consiste en el diseño de una metodología de inferencia de conocimiento: ya sean subconjuntos de genes con igual comportamiento o redes de asociación de genes.

12.2.5. Análisis de Relevancia Biológica

Finalmente, tras cualquier análisis de alto nivel es necesario evaluar la relevancia biológica del modelo resultado, es decir, interpretar los datos según un cálculo de significancia. Esta tarea no suele aparecer como objetivo dentro de las etapas a realizar en un experimento de microarray y no se menciona en la literatura, como parte de los experimentos de microarray, hasta la contribución de [2] en 2006.

Una vez que el análisis de alto nivel nos proporciona un modelo, ya sea un *ranking* de genes o un grafo de asociación de genes, dicho modelo puede verse como una lista de genes. La tarea de evaluar la relevancia biológica consiste en buscar esa lista de genes en repositorios de información biológica real. Existe un gran número de repositorios o bases de datos públicas donde los genes están anotados por su funcionalidad, entre ellos destaca: KEGG, Gene Ontology, GenBank, UniGene, Entrez Gene, etc. Buscar la lista de genes en esos repositorios no consiste en una simple úsqueda entre la lista y el repositorio. Por ejemplo, en KEGG los genes se encuentran anotados por grupos según la ruta biológica en la que actúe y en GO los genes se agrupan según categórías o términos formando una ontología. Debido a que dichos repositorios no son meras listas de genes, la tarea consiste en calcular el *match* de la lista de genes en cada una de las rutas o términos y además proporcionar un valor de significancia de ese *match*. Esta tarea requiere de cálculos estadísticos como son el test de Fisher y cálculo de correcciones sobre el mismo, mediante métodos como el de Bonferroni o Westfall.

Figura 12.3: Los autores en la anterior "etapa laboral" en la que también fueron compañeros además de hermanos.

12.3. A modo de conclusión

Para concluir queremos agradecer la invitación a este evento así como la satisfacción que nos hace el poder participar.

Isabel: cuando éramos niños soñábamos con estar en un hospital o en un campo de fútbol (véase figura 12.3), mi hermano creo que no acertó mucho y yo me acerqué algo trabajando en el tema de Bioinformatica. Pienso que esto se lo debemos a mis padres. Gracias Papá por tantos consejos, que aunque nunca fueron dados de manera directa, han resultado su efecto en lo académico y lo personal. Gracias por hacernos en nuestra niñez de cuenta cuentos y dejarnos jugar en aquella pizarra de Gonzalo de Bilbao. En la adolescencia, perdona por aquellos trabajos de literatura que me corregías el día antes de la entrega y por hacerte llevarme más de un día a clase en Reina Mercedes.

Juan Antonio: siguiendo con estos de los genes, decir que no existe un gen de la Lógica y que por lo tanto no se hereda. Recuerdo una vez, hace ya bastantes años, que compartí un curso con Ignacio y Fernando (autores de un trabajo también en este volumen), así como con Enrique, y me hablaban como si yo supiera de temas lógicos. Recuerdo, entre risas, comentarles que compartía apellido con mi padre pero no sabía nada de lógicas no convencionales, primer orden, segundo orden, etc. Descartada la herencia nos queda el contagio. En el día a día, anécdota a anécdota, tiene uno un padre que contagia entusiasmo, energías y sobre todo optimismo ante todo lo relacionado con el Saber, el *mundo académico* y la Vida en general. Muchas gracias Papá y felicidades.

Bibliografía

[1] Brown, P. and Botstein, D. (1999). Exploring the new world of the genome with dna microarrays. *Nature Genetics*, **21**(1 Suppl), 33–37.

[2] Olson, N. (2006). The microarray data analysis process: from raw data to biological significance. *NeuroRx*, **3**, 373–383.