UNIVERSIDAD
DE GRANADA

Facultad de Ciencias
E.T.S. Ingenierías Informática y de Telecomunicación

Doble Grado en Ingeniería Informática y Matemáticas

TRABAJO DE FIN DE GRADO

Metodologías Multivariantes para la Identificación de Patrones Biológicos

Presentado por:
Quintín Mesa Romero

Curso académico 2024-2025

Metodologías Multivariantes para la Identificación de Patrones Biológicos

Quintín Mesa Romero

Quintín Mesa Romero *Metodologías Multivariantes para la Identificación de Patrones Biológicos*.

Trabajo de fin de Grado. Curso académico 2024-2025.

Responsable de tutorización

José Luis Romero Béjar
Departamento de Estadística e Investigación Operativa

Doble Grado en Ingeniería Informática y Matemáticas

Facultad de Ciencias
E.T.S. Ingenierías Informática y de Telecomunicación

Universidad de Granada

Declaración de originalidad

D./Dña. Quintín Mesa Romero

Declaro explícitamente que el trabajo presentado como Trabajo de Fin de Grado (TFG), correspondiente al curso académico 2024-2025, es original, entendido esto en el sentido de que no he utilizado para la elaboración del trabajo fuentes sin citarlas debidamente.

En Granada a February 22, 2025

Fdo: Quintín Mesa Romero

1 Introducción

La biología, como disciplina científica, ha experimentado una evolución notable en las últimas décadas, pasando de enfoques cualitativos y descriptivos a un análisis más detallado y cuantitativo de los organismos vivos. Este cambio de paradigma se produjo a mediados del siglo XX con la llegada de la **biología molecular**, tras casi dos siglos de preeminencia del naturalismo basado en la observación y la contemplación. Este avance marcó el inicio de una nueva era en la que el desarrollo de ciertas herramientas tecnológicas permitió analizar los diversos y complejos niveles de organización de los organismos, generando grandes volúmenes de datos en periodos relativamente cortos: las **ciencias ómicas**.

En este contexto, las ciencias ómicas surgieron como un marco integrador que engloba el conocimiento derivado de la aplicación de tecnologías avanzadas para el estudio a nivel molecular de los distintos elementos que conforman los sistemas biológicos, como células, tejidos e individuos. Estas disciplinas no solo permiten analizar la complejidad interna de los organismos, sino también comprender las interacciones dinámicas entre sus componentes internos y los factores externos con los que estos interactúan. Ofrecen, en definitiva, una perspectiva holística del individuo, proporcionando una visión detallada del funcionamiento de sus células y de la influencia del entorno que las rodea.

El término **"ómica"** fue acuñado en la década de 1980 para referirse al estudio de **conjuntos de moléculas** específicas, como genes (genómica), transcripciones de ARN (transcriptómica), proteínas (proteómica) o metabolitos (metabolómica), entre otros. Estas disciplinas han evolucionado significativamente gracias a los avances tecnológicos que permiten abordar la complejidad inherente de los sistemas biológicos analizados. De hecho, este es el máximo distintivo de las ciencias ómicas: el uso de las llamadas **"tecnologías ómicas"**, herramientas de alto rendimiento diseñadas para generar grandes cantidades de datos en un solo experimento a partir de una única muestra. Este enfoque masivo en la obtención de datos, conocido como **"Big Data"**, ha transformado profundamente el análisis biológico, permitiendo explorar dinámicas moleculares con un gran nivel de detalle.

La integración de las ciencias ómicas con metodologías avanzadas de análisis, como las técnicas multivariantes y el aprendizaje automático, ha marcado un hito en la investigación biomédica, abriendo nuevas fronteras en la comprensión de los complejos sistemas biológicos. Estas metodologías, que permiten gestionar y analizar grandes volúmenes de datos con múltiples dimensiones, son fundamentales para descubrir patrones biológicos subyacentes que, de otro modo, podrían pasar desapercibidos utilizando métodos tradicionales. Técnicas multivariantes, como el análisis de componentes principales (PCA), el análisis clúster, el análisis factorial o el análisis discriminante, facilitan la identificación de relaciones y la reducción de la dimensionalidad en los datos, lo que es crucial para poder extraer información relevante de los vastos

conjuntos de datos generados.

A medida que los volúmenes de datos generados por las tecnologías ómicas se incrementan, la **bioinformática** se ha consolidado como una disciplina esencial para procesar, gestionar y analizar dichos datos. Facilita la identificación y visualización de patrones biológicos complejos a partir de grandes bases de datos, mediante el uso de algoritmo avanzados, herramientas computacionales y modelos estadísticos. Este enfoque es fundamental para descubrir asociaciones moleculares, determinar biomarcadores relevantes y comprender las bases genéticas de enfermedades. En este sentido, las herramientas bioinformáticas, como los lenguajes de programación R y Python, junto con plataformas especializadas como Bioconductor, permiten realizar análisis profundos de datos ómicos a gran escala, proporcionando los recursos necesarios para un manejo efectivo y preciso de la información biológica.

La combinación de estas técnicas con enfoques de **aprendizaje automático** ha transformado la capacidad para identificar patrones biológicos complejos, lo que, a su vez, ha facilitado el diagnóstico temprano de enfermedades, la clasificación de subtipos de enfermedades y el diseño de terapias personalizadas. El aprendizaje automático permite la creación de modelos predictivos que, basados en datos moleculares, pueden predecir la progresión de enfermedades o identificar biomarcadores específicos, todo ello con un nivel de precisión cada vez mayor. Estas capacidades están impulsando un cambio hacia una medicina más precisa y efectiva, en la que los tratamientos se ajustan no solo al tipo de enfermedad, sino también a las características moleculares y genéticas del paciente.

Además, la combinación de las ciencias ómicas con estas metodologías avanzadas no solo ha ampliado nuestra comprensión de los procesos biológicos fundamentales, sino que también ha proporcionado herramientas clave para el desarrollo de nuevas estrategias diagnósticas y terapéuticas. Las técnicas multivariantes y el aprendizaje automático se han convertido en pilares fundamentales en la identificación de patrones biológicos relacionados con diferentes enfermedades, desde cánceres hasta enfermedades neurodegenerativas, y en la predicción de la respuesta a distintos tratamientos. Esta integración ha sentado las bases para el avance hacia la **medicina personalizada y de precisión**, donde los tratamientos se adaptan a las características individuales de cada paciente, mejorando la efectividad y reduciendo los efectos secundarios. En este contexto, la aplicación de estas metodologías avanzadas no solo representa un avance en la investigación biomédica, sino también una prometedora realidad para la práctica clínica, abriendo la puerta a nuevas oportunidades para el tratamiento y la prevención de enfermedades de una manera mucho más específica y eficiente.

En el presente trabajo, se explorará el uso de la transcriptómica, como ciencia ómica y las metodologías avanzadas de análisis de datos, como las técnicas multivariantes y el aprendizaje automático, para la identificación y clasificación de ciertos patrones bi-

ológicos. Se realizará una revisión teórica de las técnicas multivariantes más comunes, anteriormente mencionadas, aunque nos centraremos en una de ellas con el fin de proporcionar una base sólida para su aplicación práctica en datos ómicos. Posteriormente, se llevará a cabo una implementación realista y funcional para el análisis de datos biológicos, aplicando técnicas de aprendizaje automático para la identificación de patrones biológicos significativos. A través de estas metodologías avanzadas, se intentará simplificar los datos ómicos para poder extraer la información clave que permita clasificar y entender mejor los patrones biológicos, mejorando así la precisión de los modelos predictivos.

Parte I

DATOS ÓMICOS

2 Datos ómicos

A la información cuantitativa y cualitativa obtenida a partir de las tecnologías utilizadas en las distintas ciencias ómicas, se le denomina **datos ómicos**. Estos datos abarcan información genética (genómica), de expresión génica (transcriptómica), de proteínas (proteómica), metabolitos (metabolómica) y otras áreas emergentes dentro de las ciencias ómicas.

Una de sus características más relevantes es su **alta dimensionalidad**, lo que genera conjuntos de datos masivos y complejos. Esta naturaleza multidimensional y heterogénea de los datos ómicos presenta desafíos significativos en su procesamiento, análisis e interpretación.

2.1 Estructura de los datos

Los datos con los que trabajaremos se caracterizan por tener una estructura parecida. Analizaremos un conjunto con pocas muestras frente al gran número de características que observaremos sobre ellas. Apreciamos aquí el carácter de alta dimensionalidad de los datos ómicos.

Las características que analizamos pueden ser de diferentes tipos, como el nivel de fluorescencia, en el caso de que estemos trabajando con microarrays [1], como los de ADN, metilación o proteínas, o el número de lecturas alineadas obtenidas en procedimientos de secuenciación. Estas características, pueden estar asociadas a un elemento de análisis o a un conjunto de muestras en un microarray. O bien, la información puede corresponder a un gen, un exón[2], una proteína o una región específica del genoma.

Denotaremos por N al número de características observadas, que será un valor relativamente grande, del orden de miles. Como hemos mencionado anteriormente, estas características se observan sobre un conjunto reducido de individuos, del orden de las decenas, en el mejor de los casos. Sea entonces n el número de muestras sobre las que serán observadas las variables (características).

Por consiguiente, el problema se enmarca dentro del campo de la **estadística de alta dimensión**. Esta situación, donde N supera a n, contrasta con lo que se observa en los enfoques estadísticos convencionales, en los cuales suele ocurrir todo lo contrario: el número de muestras es mayor que el de variables. Aunque esta desigualdad presenta limitaciones, también abre un nuevo campo de investigación con retos que los métodos tradicionales no pueden resolver, lo que motiva el desarrollo de nuevos

[1] Microarray: La tecnología de microarrays permite estudiar la expresión de múltiples genes simultáneamente. Consiste en fijar miles de secuencias génicas en un chip de vidrio. Al exponer una muestra de ADN o ARN, el apareamiento de bases complementarias genera una señal luminosa medible, indicando los genes expresados en la muestra.

[2] Exón: un exón es una región del genoma que termina dentro de una molécula de ARN mensajero. Algunos exones son codificantes, es decir, contienen información para fabricar una proteína, mientras que otros son no codificantes. Los genes del genoma están formados por exones e intrones.

procedimientos que se explorarán más adelante.

Las características las almacenaremos en una matriz, que llamaremos **matriz de expresión**, dada por:

$$x = [x_{ij}]_{i,j=1,\dots,n}$$

donde x_{ij} cuantifica la característica i en la muestra j.

Nota. Observemos que en un contexto estadístico convencional, la matriz de datos sería la matriz transpuesta de la que vamos a estar utilizando.

En el supuesto de que x_{ij} esté asociado con un microarray de ADN, entonces, mide un nivel de fluorescencia, tomando valores positivos, aunque pudiera ser que, tras el procesado de los datos, se diera lugar a expresiones negativas. Por su parte, si se tratase de un dato obtenido mediante la técnica de secuenciación RNA-seq, tendríamos conteos; número de lecturas cortas que se alinean sobre un gen, exón o una zona genómica concretos. Un mayor número de lecturas será indicativo de una mayor expresión de dicha característica.

Los valores observados de una característica sobre el conjunto de todas las muestras (una fila de la matriz de expresión) son, en el ámbito de la transcriptómica, lo que se conoce como *perfil*, o de forma más general, perfil de expresión.

En la matriz x los valores correspondientes a las diferentes muestras son independientes entre sí, aunque pueden haber sido obtenidos bajo condiciones distintas. Por lo tanto, no se trata de réplicas de una misma condición experimental, sino de observaciones independientes. Es decir, presentan independencia condicional. Sin embargo, las filas de x representan vectores que sí están relacionados. Por ejemplo, en una matriz de expresión génica, los valores de expresión de las filas no son independientes, debido a que los genes tienden a actuar de manera coordinada.

Por lo general, los datos en las columnas de la matriz x, no pueden compararse directamente entre sí, por la presencia de diversos artefactos técnicos y ruido en la medición de la característica de interés. Es por ello que se han desarrollado técnicas para corregir estos problemas, denominadas como **técnicas de preprocesado**. Al aplicar estos métodos, los datos dejan de ser completamente independientes. No obstante, en la mayoría de los estudios este aspecto no se suele tener en cuenta. Tras la normalización, los datos siguen considerandose independientes por columnas (muestras) y dependientes por filas.

A la información o variables que describen y caracterizan a las muestras, las llamaremos **metadatos** o **variables fenotípicas**. En este contexto, el uso de este término es adecuado porque estas variables reflejan atributos medibles y observables de las muestras, lo que se conoce en el ámbito de la biología como *fenotipo*. Normalmente

tendremos varias variables fenotípicas. Llamaremos $y = (y_1, ..., y_n)$ a los valores observados de una variable en las n muestras. Uno de los casos más típicos de variable fenotípica es cuando se tienen dos grupos de muestras: casos (individuos que tienen la enfermedad) y controles (no tienen la enfermedad o condición de interés). En este caso tendríamos $y_i = 1$, para un caso e $y_i = 0$, si es control. Si tuvieramos la situación en la que hay k grupos a comparar, con $k > 2$, entonces se utiliza $y_i \in \{1, ..., n\}$ con $i = 1, ..., n$. Hemos de recalcar que los valores y_k son arbitrarios y pueden tomar cualquier otro par de valores.

2.2 Problema Estadístico

Normalmente, las técnicas estadísticas utilizadas en muchos campos se basan en contextos en los que el número de muestras, n, es mayor que el de variables, N. Sin embargo, en el caso de los datos ómicos, esta relación se invierte, lo que obliga a ajustar estos procedimientos de manera que, en algunos casos, la adaptación resulta más o menos exitosa. En otras palabras, la falta de suficientes muestras para la cantidad de variables presentes, hace que sea extremadamente difícil encontrar un modelo que pueda capturar de manera precisa la relación entre las variables predictores y la variable respuesta. Esto se debe a que no tenemos una cantidad adecuada de datos para entrenar de manera efectiva un modelo estadístico que pueda generalizarse de manera fiable a nuevas observaciones.

La dificultad de analizar datos de alta dimensionalidad resulta además de la conjunción de dos efectos.

En primer lugar, los espacios de alta dimensión tienen propiedades geométricas que son contra-intuitivas y alejadas de las propiedades que se pueden observar en espacios bidimensionales o tridimensionales.

En segundo lugar, las herramientas de análisis de datos suelen diseñarse teniendo en cuenta propiedades intuitivas y ejemplos en espacios de baja dimensión; por lo general, las herramientas de análisis de datos se ilustran mejor en espacios de dos y tres dimensiones, por razones obvias. El problema es que esas herramientas también se utilizan cuando los datos son de alta dimensión y más complejos. En este tipo de situaciones, perdemos la intuición del comportamiento de las herramientas y podemos sacar conclusiones erróneas sobre sus resultados, dificultando la construcción de modelos estadísticos precisos.

Por todo ello, en este contexto, las técnicas estadísticas utilizadas son mera aplicación de procedimientos diseñados para la situación antes comentada en la que el número de muestras es mayor que el de variables.

Uno de los principales retos que se abordarán es el análisis de expresión diferencial, que examina cómo varían los niveles de expresión génica[5] entre distintas condiciones

[5]Expresión génica: La expresión génica es el proceso por el cual la información codificada por un gen se usa para producir moléculas de ARN que codifican para proteínas o para producir moléculas

experimentales o grupos de individuos. En particular, se busca determinar si existe una relación entre el perfil de expresión génica y una variable fenotípica específica. Este enfoque, denominado análisis de expresión diferencial marginal, permite explorar asociaciones entre conjuntos de genes, organizados como grupos de filas dentro de la matriz de expresión, y la característica fenotípica de interés. A este tipo de análisis se le conoce también como análisis de conjuntos de genes o *gene set analysis*, y su objetivo es identificar patrones de expresión que puedan estar vinculados a determinados rasgos biológicos.

2.3 Transcriptómica: datos RNA-seq y single-cell RNA-seq

Entre las distintas tecnologías utilizadas en la generación de datos ómicos, la **transcriptómica** desempeña un papel fundamental en el análisis de la expresión génica, y en particular, las tecnologías de RNA-seq y single-cell RNA-seq han revolucionado este campo al permitir la cuantificación precisa de los niveles de ARN mensajero en diferentes condiciones biológicas. Estas técnicas producen datos con una estructura matricial compleja, caracterizada por un alto número de variables (genes) frente a un número reducido de muestras (individuos o células). Esta estructura plantea desafíos en términos de almacenamiento, procesamiento y análisis, debido a la alta dimensionalidad de los datos generados. Nos centraremos en la transcriptómica por su capacidad para ofrecer una visión dinámica y profunda de la actividad celular, permitiendo identificar patrones de expresión génica que reflejan procesos biológicos clave.

En este apartado, se describirá la estructura de los datos obtenidos mediante RNA-seq y single-cell RNA-seq, y se discutirán los principales retos asociados a su manejo, desde las consideraciones técnicas hasta las implicaciones estadísticas y computacionales, que hacen de la transcriptómica un área idónea para aplicar metodologías multivariantes en la identificación de patrones biológicos.

¿Qué es la transcriptómica? Tecnologias

La transcriptómica es la rama de la biología que estudia el conjunto completo de ARN (ácido ribonucleico) transcritos en una célula, tejido u organismo en un momento determinado, bajo condiciones específicas. A este conjunto se le denomina transcriptoma. Se centra en la cuantificación y caracterización de los distintos tipos de ARN, incluyendo ARN mensajero (mARN), ARN de transferencia (tRNA), ARN ribosomal (rRNA) y ARN no codificante (ncRNA), entre otros. Este campo ha evolucionado significativamente desde la formulación del dogma central de la biología molecular por Francis Crick en 1958, que estableció la transferencia de información genética desde el

de ARN no codificantes que cumplen otras funciones. La expresión génica actúa como un "interruptor" que controla cuándo y dónde se producen moléculas de ARN y proteínas y como un "control de volumen" para determinar qué cantidad de esos materiales se produce

ADN al ARN y posteriomennte a las proteínas.

A medida que la transcriptómica ha avanzado, se han ido desarrollando varias tecnologías para deducir y cuantificar el transcriptoma, basadas tanto en hibridación como en secuenciación. Los enfoques basados en hibridación, como los microarrays, pese a que son más económicos y tienen un alto rendimiento, dependen del conocimiento existente sobre la sencuencia del genoma.

A diferencia de los métodos basados en microarrays, los enfoques basados en secuencias determinan directamente la secuencia de ARN. Inicialmente, se utilizó la secuenciación de Sanger de bibliotecas de ARN, pero era bastante costosa y de bajo rendimiento y generalmente no cuantitativa. Se desarrollaron métodos basados en etiquetas para superar estas limitaciones, pero tenían el inconveniente de que estaban basados en la costosa tecnolgía de secuenciación de Sanger, y una parte significativa de las etiquetas cortas no se podían asignar de forma única al genoma de referencia. Todas estas desventajas limitan el uso de la tecnología de secuenciación tradicional para anotar la estructura de los transcriptomas.

Tecnologías de secuenciación de ADN de alto rendimiento como **RNA-seq** y **single-cell RNA-seq** han emergido como herramientas clave para estudiar la expresión génica a gran escala. Estas técnicas permiten la cuantificación precisa de los niveles de ARN en diferentes condiciones biológicas, a diferencia de los métodos basados en microarrays, lo que proporciona información valiosa sobre la actividad celular y los mecanismos de regulación genética. Sin embargo, los datos obtenidos mediante RNA-seq y single-cell RNA-seq poseen características específicas que influyen en su representación y análisis. Estas tecnologías generan grandes volúmenes de datos con una estructura matricial compleja, en la que el número de características (genes) supera ampliamente al número de muestras, lo que da lugar a retos significativos en términos de almacenamiento, procesamiento y análisis.

RNA-seq

El método RNA-seq (secuenciación de ARN) consiste en la conversión de una muestra de ARN (total o fraccionado) en una biblioteca de ADNc (ADN codificado), que luego es secuenciada utilizando tecnologías de secuenciación profunda. Genera un conjunto masivo de datos que consiste en lecturas cortas de ARN transcrito, secuencias que generalmente varían entre 30 y 400 pares de bases de longitud, representando fragmentos de transcritos provenientes de ARN mensajero (ARNm) o ARN no codificante. Estas secuencias se alinean con un genoma de referencia o con transcritos de referencia para mapear la estructura transcripcional y cuantificar la expresión génica. Una de las principales aplicaciones de RNA-seq es el análisis de expresión diferencial, mencionado en la sección previa y que abordaremos de forma práctica, que permite comparar los niveles de expresión de genes entre diferentes condiciones biológicas, como células tratadas frente a no tratadas, tejidos sanos frente a cancerosos o distintos estados del

desarrollo. Esto ofrece una visión detallada de los cambios en la actividad génica y ayuda a identificar biomarcadores, rutas metabólicas alteradas o procesos reguladores clave. Además, RNA-seq no está limitado a detectar solo transcritos que corresponden a secuencias genómicas conocidas, lo que lo hace particularmente útil para organismos no modelo o cuando se carece de un genoma de referencia bien caracterizado.

En la secuenciación de ARN, las lecturas generadas a partir de las muestras de ARN se alinean contra un genoma de referencia o se ensamblan de nuevo para crear un "mapa transcripcional". Si se dispone de un genoma de referencia, los datos se alinean para identificar la ubicación exacta de los transcritos en el genoma, permitiendo la cuantificación de la expresión génica. En casos donde no hay un genoma de referencia, las lecturas de ARN se ensamblan para generar una secuencia de contigs [3] que luego se pueden anotar funcionalmente. Además de mapear las lecturas a un genoma, se deben identificar eventos de empalme (splicing [4]), que es crucial para detectar variantes de splicing alternativo. Este proceso es especialmente importante para genes que tienen varios exones, ya que las lecturas pueden cruzar estos empalmes y revelar alternativas de splicing que no son evidentes con tecnologías anteriores.

Pese a las ventajas que la RNA-seq tiene frente a tecnologías anteriores, los conjuntos de datos producidos son grandes y complejos y la interpretación no es sencilla. La interpretación de los datos de secuenciación de ARN depende de la cuestión científica de interés. El objetivo principal de muchos estudios biológicos es el perfil de expresión génica entre muestras, que es particularmente relevante, por ejemplo, para experimentos controlados que comparan la expresión en cepas de tipo salvaje y mutantes del mismo tejido, comparando células tratadas versus no tratadas, cáncer versus normal, etc.

Por otra parte, los datos RNA-seq requieren estar en unos formatos específicos para su tratamiento. Formatos adecuados para almacenar secuencias tanto de ácidos nucleicos como de proteínas. Estos son: formato **FASTA** y **FASTQ**.

- **Formato FASTA:**

- **Formato FASTAQ:** es el más popular y consiste en cuatro líneas por lectura:
  - La primera comienza con el carácter "@" y contiene el nombre de la secuencia. Opcionalmente, puede incluir una descripción.
  - La segunda línea contiene la secuencia con las letras correspondientes, dependiendo del tipo de secuencia del que se trate (nucleótido o aminoácido).
  - La tercera comienza con el carácter "+" y contiene información opcional sobre la secuencia.

[3]Contig: Tramo de secuencia continua in silico generada por alineamiento de lecturas de secuencias solapantes.

[4]Splicing: el splicing o empalme, ocurre al final del proceso de transcripción e implica cortar y reorganizar secciones de ARNm.

– La cuarta y última línea cuantifica la calidad o confiabilidad de cada base en la secuencia recogida en la segunda línea, basada en el índice *Phred* y su codificación.

sigle-cell RNA-seq

Parte II

FUNDAMENTOS MATEMÁTICOS

Parte III

FUNDAMENTOS INFORMÁTICOS