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 20242025

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 20242025.

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 20242025, es original, entendido esto en el sentido de que no he utilizado para la elaboración del trabajo fuentes sin citarlas debidamente.

En Granada a March 10, 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: la *era de 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 voluminosos 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 algoritmos 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, entre otros, 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 biológicos. 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.

En este capítulo, se presenta la estructura de los datos ómicos, destacando su naturaleza matricial en la que el número de características (variables) suele superar ampliamente el número de muestras. Se analiza el desafío estadístico que representan los datos de alta dimensión y se introduce el problema de la expresión diferencial. Además, se aborda la transcriptómica y las tecnologías de secuenciación de ARN (RNA-seq y scRNA-seq), describiendo la estructura de los datos que generan y los retos asociados a su manejo, dado su gran volumen y alta dimensionalidad.

La información presentada en las próximas dos secciones ha sido extraída de la fuente bibliográfica[1].

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.

[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].

[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

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 procedimientos que se explorarán más adelante.

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

$$x = [x_{ij}]_{i=1,\dots,N,\, j=1,\dots,n} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nn} \end{bmatrix}$$

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, que será introducida a continuación, 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 indepen-

que otros son no codificantes. Los genes del genoma están formados por exones e intrones, que son trozos muy grandes de ARN dentro de una molécula de ARN mensajero que interfieren con el código de los exones. Estos intrones se eliminan de la molécula de ARN para dejar una serie de exones unidos entre sí de manera que se puedan codificar los aminoácidos correctos[3][4]

dientes 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[5], 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*[6]. 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

[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 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[5].

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 entre distintas condiciones 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.

2.3.1 ¿Qué es la transcriptómica? Tecnologias de secuenciación

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 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[7].

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.

2.3.2 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[8].

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[11].

Pese a las ventajas que la RNA-seq tiene frente a tecnologías anteriores, los conjun-

[3]Contig: Tramo de secuencia continua in silico generada por alineamiento de lecturas de secuencias solapantes[9].

[4]Splicing: el splicing o empalme, ocurre al final del proceso de transcripción e implica cortar y reorganizar secciones de ARNm[10].

tos 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[12].

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*. La información que presentamos a continuación ha sido extraída de[1].

- **Formato FASTA:** basado en texto, es usado para representar secuencias de nucleótidos o de aminoácidos, ambos representados mediante una sola letra. Incluye símbolos para representar huecos (*gaps*) o posiciones desconocidas en la secuencia. Consta de dos líneas:
  - La primera línea comienza con el símbolo >, junto con una descripción de la secuencia.
  - La segunda, contiene la secuencia de bases o aminoácidos.

  Pese a que no hay restricciones en el número de filas, el número de columnas no debería superar las 80.
- **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.
  - 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.

Ejemplo

```
@SRR1293399.1 ILLUMINA-545855_0026_FC629BG:6:1:1022:5049 length=50
ACAGGGACGCCATCGAATCCGGATCNTNNNNNNNNNNNANNNNNNNNNN
+SRR1293399.1 ILLUMINA-545855_0026_FC629BG:6:1:1022:5049 length=50
dee\edYcdc`bbY`S]bb_]]Ua^BBBBBBBBBBBBBBBBBBBBBBBB
```

2.3.3 scRNA-seq (sigle-cell RNA-seq)

La secuenciación de ARN ha impulsado muchos descubrimientos e innovaciones en diversos campos en los últimos años. Por razones prácticas, la técnica RNA-seq suele realizarse en muestras que comprenden entre miles y millones de células. Sin embargo, esto ha dificultado la evaluación directa de la unidad fundamental de la biología: la célula. Es por esto que surge una variante a la RNA-seq: *scRNA-seq* (*single-cell RNA-seq*), que se ha extendido considerablemente hasta consolidarse como la opción principal en investigación de la diversidad celular, ya que posibilita el estudio individual de cada célula dentro de una misma muestra. Es una tecnología que permite la cuantificación y comparación de los transcriptomas de células individuales, logrando una disección de la expresión génica a resolución de una sola célula y describiendo moléculas de ARN con alta precisión y a escala genómica. Además, otro aspecto importante de esta técnica es que permite analizar la heterogeneidad celuar; evaluar las similitudes y diferencias transcripcionales dentro de una población de células, lo que ha permitido una comprensión más detallada de los procesos moleculares al evidenciar la variabilidad existente dentro de una misma población celular.

Aunque existe una gran confianza en la utilidad general de scRNA-seq, hay una barrera técnica que debe considerarse cuidadosamente: el aislamiento efectivo de células individuales del tejido de interés, pues existe un riesgo potencial de que los protocolos utilizados para este proceso alteren los niveles de ARNm. Aunque las células vecinas pueden contribuir a mantener los estados celulares, scRNA-seq opera bajo el supuesto de que el aislamiento de células individuales no desencadena cambios transcriptómicos antes de la captura del ARNm. Además, esta técnica a menudo requiere el uso de marcadores específicos para distinguir con precisión diferentes poblaciones celulares, lo que puede añadir complejidad al diseño experimental. A esto se suman limitaciones relacionadas con el tiempo y los costos asociados, que pueden dificultar la realización de investigaciones a gran escala.

El análisis de datos provenientes de scRNA-seq supone un desafío computacional considerable debido a su alta dimensionalidad y a la presencia de ruido significativo. En el estudio de muestras heterogéneas, es fundamental identificar las distintas subpoblaciones celulares presentes, lo que requiere el desarrollo de algoritmos capaces de descomponer la mezcla de expresión génica obtenida. Para ello, se aplican técnicas de normalización y reducción de dimensionalidad, como el análisis de componentes principales (PCA), entre otras, que permiten segmentar las células en función de sus perfiles de expresión. No obstante, la presencia de efectos por lote y la variabilidad técnica pueden introducir sesgos en los resultados, lo que hace necesario implementar estrategias de corrección para garantizar la robustez del análisis y evitar la aparición de artefactos en la identificación de patrones biológicos.

Además, el volumen de datos generado por scRNA-seq plantea retos en términos de almacenamiento, gestión y escalabilidad computacional. La integración de millones

de lecturas individuales requiere estrategias eficientes de compresión y procesamiento distribuido, así como el uso de infraestructuras de alto rendimiento para análisis en estudios de gran escala. Dado que el aislamiento manual de células en el laboratorio puede ser costoso y técnicamente desafiante, se han propuesto enfoques computacionales que permitan inferir la composición celular sin necesidad de manipulación experimental. En este contexto, los métodos multivariantes desempeñan un papel clave en la extracción de información relevante, facilitando la identificación de estructuras en los datos sin necesidad de un conocimiento previo sobre los tipos celulares presentes en la muestra. La combinación de técnicas estadísticas y aprendizaje automático ha demostrado ser esencial para mejorar la precisión y fiabilidad de estos análisis en el ámbito biomolecular.

Dado que scRNA-seq es una variante de RNA-seq, hereda sus formatos de almacenamiento iniciales, principalmente FASTA y FASTQ, siendo este último el formato más crudo en el que se encuentran los datos de scRNA-seq.

Toda la información recogida en esta subsección ha sido extraída de las fuentes bibliográficas[13],[14],[15].

Parte II

FUNDAMENTOS MATEMÁTICOS

3 Técnicas multivariantes: fundamentos y desarrollo del análisis clúster

El análisis de datos en ciencias ómicas requiere metodologías capaces de manejar la complejidad inherente a los sistemas estudiados. En particular, las técnicas multivariantes han demostrado ser herramientas fundamentales para la exploración, modelado e interpretación de datos de alta dimensión. Estas metodologías permiten identificar relaciones entre variables, reducir la dimensionalidad y clasificar observaciones en función de patrones subyacentes.

Este capítulo está estructurado en dos secciones. En primer lugar, se presentarán las principales técnicas multivariantes, destacando su utilidad y objetivos dentro del análisis de datos. Posteriormente, se abordará en profundidad el análisis clúster, una técnica multivariante ampliamente utilizada para la identificación de patrones en grandes volúmenes de datos. Su aplicación en el ámbito biológico permite revelar estructuras subyacentes en datos complejos, facilitando la comprensión de procesos como la agrupación de expresiones génicas o la clasificación de organismos en función de sus características.

Dado que en el capítulo dedicado a los datos ómicos hemos introducido la matriz de datos ómicos X, que representa las N características medidas sobre n muestras, mantendremos esta notación en el desarrollo de los fundamentos matemáticos sobre los que se basa este trabajo. Así, consideraremos que la matriz de expresión $X \in \mathbb{R}^{N\times n}$ almacena las observaciones de nuestras variables, con filas representando las características y columnas las muestras.

3.1 Preliminares

La información aquí recogida se ha extraído principalmente de las fuentes [16], [17], [18].

El análisis multivariante es una herramienta clave para explorar y comprender la complejidad de los sistemas biológicos, económicos y sociales. Su capacidad para procesar múltiples variables simultáneamente permite identificar patrones ocultos en grandes volúmenes de datos.

Las técnicas multivariantes son fundamentales para abordar la complejidad de los datos en diversas disciplinas, incluyendo las ciencias ómicas, donde se requieren metodologías capaces de gestionar la alta dimensionalidad y variabilidad de los datos

obtenidos. Estas herramientas permiten descubrir relaciones entre variables, reducir la dimensionalidad y clasificar observaciones, lo que facilita la interpretación y el modelado de sistemas complejos.

El desarrollo del análisis multivariante se remonta a principios del siglo XX, cuando pioneros como Karl Pearson y R.A Fisher introdujeron técnicas fundamentales como el análisis de componentes principales y el análisis discriminante. Posteriormente, C.R. Rao y otros investigadores expandieron estos métodos, estableciendo bases matemáticas sólidas que han permitido su aplicación en un amplio espectro de disciplinas. Estas técnicas han ido desarrollándose exponencialmente con el avance de la computación, facilitando el procesamiento de grandes volúmenes de datos y dando lugar a análisis mucho más sofisticados en muchas áreas como la biología, la economía, las ciencias sociales, etc.

En términos generales, las metodologías multivariantes pueden dividirse en dos grandes enfoques: *descriptivo* e *inferencial*. El primero busca simplificar la estructura de los datos y revelar relaciones latentes entre variables, mientras que el segundo permite realizar pruebas de hipótesis considerando múltiples variables de manera simultánea, asegurando la validez estadística de los resultados. La elección de la técnica adecuada depende del tipo de datos y de la pregunta de investigación. A continuación, se presentan algunas de las principales metodologías multivariantes:

3.1.1 Análisis de Componentes Principales (PCA)

El *análisis de componentes principales (PCA)* fue introducido por primera vez por Karl Pearson a principios del siglo XX. El tratamiento formal de esta técnica se debe a Hotelling (1933) y Rao (1964). Su propósito era facilitar la comprensión de conjuntos de datos complejos mediante la reducción de su dimensionalidad, minimizando la pérdida de información. En PCA, un conjunto de N variables correlacionadas se transforma en un conjunto más pequeño de constructos hipotéticos no correlacionados llamados *componentes principales*. Su objetivo es condensar la información proporcionada por dichas variables en unas pocas de ellas o en pocas combinaciones lineales de ellas (con máxima variabilidad).

Las componentes principales se definen como combinaciones lineales de las variables originales que capturan la mayor variabilidad posible en los datos. Matemáticamente, si Y es un vector de N variables observadas con media μ y matriz de covarianza Σ, las componentes principales Z_i se obtienen como:

$$Z_i = p_i^{'} Y, \;\; i = 1, 2, ..., N$$

donde p_i es un vector de pesos o *cargas principales* que maximizan la varianza de Z_i bajo la restricción de que p_i tiene norma unitaria, es decir,

$$\max Var(Z_i) = p_i^{'} \Sigma p_i, \text{ sujeto a } p_i^{'} p_i = 1.$$

y tal que asegura que las componentes principales son ortogonales entre sí, es decir:

$$p_i^{'} p_j = 0, \text{ para } i \neq j.$$

Así, garantizamos que las componentes principales Z_i y Z_j son intercorrelacionadas, es decir, su covarianza es cero para $i \neq j$.

Los vectores p_i son los autovectors de la matriz de covarianza Σ, y los valores propios λ_i, corresponden a la varianza explicada por cada componente principal. La transformación completa de los datos se expresa de la siguiente forma:

$$Z = P^{'} Y$$

donde P es la matriz de autovectores de Σ, lo que garantiza que las componentes principales sean ortogonales entre sí y no correlacionadas, cada una con las anteriores.

Las CP se utilizan para descubrir e interpretar las dependencias que existen entre las variables y para examinar las relaciones que pueden existir entre los individuos. También son útiles para estabilizar las estimaciones, evaluar la normalidad multivariante y detectar valores atípicos.

3.1.2 Análisis factorial

El objetivo principal del *análisis factorial* (AF) es capturar la realidad de la manera más simple posible, identificando unas pocas variables latentes[8] que definen esa realidad. Esta técnica multivariante busca explicar el comportamiento de las N variables en la matriz de datos X utilizando un número reducido de variables latentes, denominadas *factores*. Lo ideal es que toda la información contenida en X pueda ser representada mediante un número menor de factores. Esta técnica busca explicar las correlaciones entre las variables mediante la combinación lineal de dichos factores. Así, cada factor es una variable latente que influye en las variables observadas, y cuya presencia se infiere a partir de las correlaciones entre ellas.

Matemáticamente, cada variable observada, $x \in \mathbb{R}^N$, se expresa como una combinación lineal de estos factores, más un término de error específico:

$$x_i = \sum_{l=1}^{k} q_{il} f_l + \mu_i, \ i = 1, ..., N.$$

Aquí, f_l, con $l = 1, ..., k$ denota a los factores. El número de factores, k, debería ser siempre mucho más pequeño que N.

En definitiva, el modelo de análisis factorial asume que la variable observada x puede descomponerse en dos componentes: una parte explicada por los factores co-

[8] Variable latente: variable no observable que se infiere a partir de un conjunto de variables observables utilizando un modelo matemático.

munes y una parte específica de cada variable. Esto se expresa matricialmente de la siguiente forma:

$$x = \Lambda F + \psi$$

donde:

- Λ es la matriz de cargas factoriales de dimensión $N \times k$,
- F es el vector de factores de dimensión k.
- ψ es el vector de factores específicos o residuales de dimensión N

El análisis factorial fue desarrollado por Charles Spearman a principios del siglo XX para modelar la inteligencia humana, postulando que las puntuaciones en distintas pruebas estaban intercorrelacionadas debido a un único factor latente de inteligencia general (g). Su modelo de un solo factor fue posteriormente generalizado por Thurstone a múltiples factores.

El análisis de componentes principales (PCA) y el análisis factorial suelen confundirse porque ambos analizan la variación en un conjunto de variables a partir de la matriz de correlación o covarianza. Sin embargo, mientras que en el EFA unas pocas variables latentes explican las correlaciones observadas, en el PCA se necesitan todos los componentes principales para describir completamente la variabilidad. Así, el PCA se centra en explicar la varianza total, mientras que el EFA se enfoca en las relaciones entre las variables mediante factores comunes.

3.1.3 Análisis Discriminante

El *análisis discriminante* es una técnica multivariante exploratoria que permite identificar un subconjunto de variables y funciones asociadas que maximicen la separación entre los grupos o poblaciones de estudio. Su objetivo principal es construir funciones discriminantes que describan y caractericen la separación de los grupos, evaluar el grado de diferenciación y analizar la contribución de cada variable a la discriminación.

Cuando estas funciones son combinaciones lineales de las variables originales, se denominan funciones discriminantes lineales (LDF). En particular, el análisis discriminante lineal de Fisher busca encontrar una combinación lineal de variables que maximice la separación entre grupos. Para dos grupos con medias μ_1 y μ_2 y una matriiz de covarianza común Σ, la función discriminante de Fisher se define como:

$$L = a^{'} y = \sum_{j=1}^{N} a_j y_j$$

donde a es el vector de coeficientes de la función discriminante e y es el vector de observaciones de las variables.

El vector a que maximiza la separación entre los grupos se obtiene como:

$$a_\Sigma = \Sigma^{-1}(\mu_1 - \mu_2)$$

Además, la *distancia de Mahalanobis*, que explicaremos con detalle en la siguiente sección, se emplea para medir la separación entre los centroides de los grupos:

$$D^2 = (\mu_1 - \mu_2)' S^{-1}(\mu_1 - \mu_2)$$

Si D^2 es significativo, implica una buena discriminación entre los grupos.

Este análisis tiene aplicaciones en diversos campos: en biología, Fisher (1936) lo utilizó para diferenciar especies de iris en función de características morfológicas; en la gestión de personal, permite clasificar profesionales según sus habilidades; en medicina, ayuda a distinguir entre individuos con alto o bajo riesgo de enfermedades; y en la industria, contribuye a identificar cuándo un proceso está bajo control o fuera de control.

Para el caso de múltiples grupos, la función discriminante se construye de manera que maximice la variabilidad entre los grupos en relación con la variabilidad dentro de los grupos, lo que se logra mediante una descomposición en valores propios.

Toda la información ha sido extraída de las fuentes bibliográficas[19],[20][21],[22]

3.2 Análisis Clúster

Parte III

FUNDAMENTOS INFORMÁTICOS

Bibliography

[1] Guillermo Ayala. *Bioinformática Estadística*. Edición digital en PDF. Valencia, España, 2023.

[2] National Human Genome Research Institute. *Tecnología de microarrays (chips de ADN o ARN)*. Último acceso: 23 de febrero de 2025. 2025. URL: https://www.genome.gov/es/genetics-glossary/Tecnolog%C3%ADa-de-microarrays-chips-de-ADN-o-ARN.

[3] National Human Genome Research Institute. *Exón*. Último acceso: 23 de febrero de 2025. 2025. URL: https://www.genome.gov/genetics-glossary/Exon.

[4] National Human Genome Research Institute. *Intrón*. Último acceso: 23 de febrero de 2025. 2025. URL: https://www.genome.gov/es/genetics-glossary/Intron.

[5] National Human Genome Research Institute. *Expresión génica*. Último acceso: 23 de febrero de 2025. 2025. URL: https://www.genome.gov/es/genetics-glossary/Expresion-genica.

[6] National Human Genome Research Institute. *Fenotipo*. Último acceso: 23 de febrero de 2025. 2025. URL: https://www.genome.gov/es/genetics-glossary/Fenotipo.

[7] ZhiCheng Dong and Yan Chen. *Transcriptomics: Advances and Approaches*. Recibido el 14 de agosto de 2013; aceptado el 6 de septiembre de 2013. Guangzhou 510650, China, 2013.

[8] Mark Gerstein Zhong Wang and Michael Snyder. *RNA-Seq: a revolutionary tool for transcriptomics*. Publicado en versión final editada en enero de 2009. Department of Molecular, Cellular et al., 2009. DOI: 10.1038/nrg2484.

[9] Instituto Roche. *Contig*. Último acceso: 23 de febrero de 2025. 2025. URL: https://www.institutoroche.es/recursos/glosario/contig.

[10] Your Genome. *What is RNA Splicing?* Último acceso: 23 de febrero de 2025. 2025. URL: https://www.yourgenome.org/theme/what-is-rna-splicing/.

[11] Ali Mortazavi et al. *Mapping and quantifying mammalian transcriptomes by RNA-Seq*. © 2008 Nature Publishing Group. 2008. URL: http://www.nature.com/naturemethods.

[12] Mark D. Robinson Alicia Oshlack and Matthew D. Young. *From RNA-seq reads to differential expression results*. REVIEW. 2010. URL: http://genomebiology.com/2010/11/12/220.

[13] Ashraful Haque et al. *A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications*. REVIEW, Open Access. 2017. DOI: 10.1186/s13073-017-0467-4.

[14] Dragomirka Jovic et al. *Single-cell RNA sequencing technologies and applications: A brief overview*. REVIEW, Recibido: 5 de agosto de 2021; Revisado: 9 de diciembre de 2021; Aceptado: 20 de diciembre de 2021. 2021. DOI: 10.1002/ctm2.694.

[15] Broad Institute. *Processing Single-Cell RNA-Seq Data*. Último acceso: 23 de febrero de 2025. 2019. URL: https://broadinstitute.github.io/2019_scWorkshop/processing-scrnaseq-data.html.

[16] C. Radhakrishna Rao. *Multivariate Analysis: Some Reminiscences on Its Origin and Development*. 1983. URL: https://www.jstor.org/stable/25052296.

[17] Mushtak A.K. Shiker. *Multivariate Statistical Analysis*. 2012. URL: https://www.britishjournalofscience.com.

[18] Alvin C. Rencher and William F. Christensen. *Methods of Multivariate Analysis*. 3rd. Wiley Series in Probability and Statistics. Editors: David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Harvey Goldstein, Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg. Provo, Utah: John Wiley Sons, Inc., 2012.

[19] José Luis Romero Béjar and Carlos Francisco Salto Díaz. *Tema 3.- Análisis de componentes principales (ACP)*. Asignatura: Estadística Multivariante (Prácticas). Grados en: Física y Matemáticas; Ingeniería Informática y Matemáticas; Matemáticas (4º Curso - 1er semestre 2023-2024). 2023. URL: https://digibug.ugr.es/bitstream/handle/10481/85857/ACP.pdf?sequence=1&isAllowed=y.

[20] José Luis Romero Béjar and Carlos Francisco Salto Díaz. *Tema 4.- Análisis factorial (AF)*. Asignatura: Estadística Multivariante (Prácticas). Grados en: Física y Matemáticas; Ingeniería Informática y Matemáticas; Matemáticas (4º Curso - 1er semestre 2023-2024). 2023. URL: https://digibug.ugr.es/bitstream/handle/10481/85859/AF.pdf?sequence=1&isAllowed=y.

[21] Neil H. Timm. *Applied Multivariate Analysis*. Primera edición. 2002.

[22] Wolfgang Karl Härdle and Léopold Simar. *Applied Multivariate Statistical Analysis*. Library of Congress Control Number: 2015933294. 2015. DOI: 10.1007/978-3-662-45171-7.