Práctica: Análisis de Componentes Principales (PCA) con descriptores moleculares y actividad biológica

Esta práctica está diseñada para que el estudiante de QFB pueda explorar datos de fármacos simulados usando Análisis de Componentes Principales (PCA) en RStudio, conectando los descriptores moleculares con la actividad biológica (Activa: Sí/No) generada mediante una “regla oculta” similar a un modelo QSAR.

Objetivos: Comprender qué es PCA y para qué se utiliza en el contexto de descriptores moleculares. Aplicar PCA a un conjunto de fármacos simulados con descriptores fisicoquímicos. Visualizar la distribución de compuestos en el espacio PC1–PC2 y relacionarla con la actividad biológica. Interpretar las cargas (loadings) de los componentes principales. Relacionar PC1 con la “regla oculta” utilizada para generar la actividad biológica.

Descripción general del flujo de trabajo

El código asociado a esta práctica (almacenado en el repositorio de GitHub del curso) realiza los siguientes pasos:

1. Generación del dataset de fármacos:
- Se simulan descriptores moleculares típicos: MW, LogP, TPSA, HBD, HBA.
- Se define una combinación lineal de estos descriptores (la “regla oculta”) que modela la probabilidad de actividad biológica.
- Se aplica una función logística para convertir esa combinación en una probabilidad entre 0 y 1.
- A partir de esa probabilidad se genera una variable binaria de actividad (Activa: Sí/No).

2. Preparación de los datos para PCA:
- Se toman únicamente las columnas numéricas correspondientes a los descriptores.
- Se estandarizan con la función scale(), de modo que cada descriptor tenga media 0 y desviación estándar 1 (equivalente a z-scores).

3. Cálculo del PCA:
- Se utiliza la función prcomp() de R para obtener los componentes principales.
- Se solicita un resumen con summary() para ver cuánta varianza explica PC1, PC2, etc.

4. Visualización de resultados:
- Se genera un gráfico de PC1 vs PC2 en el que cada punto representa un fármaco.
- Los puntos se colorean de acuerdo con la actividad (Activa Sí/No), de forma que se pueda observar si los compuestos activos tienden a agruparse en alguna región del plano PC1–PC2.
- Se pueden añadir etiquetas, leyendas y títulos descriptivos para facilitar la interpretación en clase.

5. Interpretación de las cargas (loadings):
- Se revisa el objeto pca$rotation para ver las cargas de cada descriptor en PC1 y PC2.
- Se interpreta qué variables contribuyen más a PC1 (por ejemplo, tamaño y LogP) y cuáles a PC2 (por ejemplo, polaridad y capacidad de formar puentes de hidrógeno).

6. Conexión con la “regla oculta”:
- Se recupera o recalcula el score_lineal utilizado en la simulación.
- Se calcula la correlación entre PC1 y dicho score_lineal.
- Si la correlación es alta, se discute en clase cómo PC1 está capturando, de forma no explícita, la misma combinación de variables que se usó para generar la actividad.

Sección 1: Generación del dataset y la regla oculta

En esta parte del script se generan valores simulados para cada descriptor molecular: - MW: peso molecular aproximado de los compuestos, con una distribución normal centrada en un valor típico para fármacos (por ejemplo, 300 Da). - LogP: coeficiente de partición octanol/agua, que refleja la lipofilicidad. - TPSA: área polar superficial topológica, relacionada con la polaridad y la capacidad de formar puentes de hidrógeno. - HBD: número de donadores de hidrógeno. - HBA: número de aceptores de hidrógeno. Luego se define una expresión lineal (score_lineal) que combina estos descriptores con un peso específico para cada uno. Este score_lineal representa una especie de “actividad intrínseca” que no es directamente observable por el modelo. La función logística (1 / (1 + exp(-score_lineal))) transforma el score en una probabilidad entre 0 y 1. A partir de esa probabilidad, se usa rbinom() para asignar a cada molécula una etiqueta de actividad (Activa = Sí o No). Esta construcción imita un modelo QSAR simple en el que ciertas propiedades (por ejemplo, LogP moderado y TPSA moderada) favorecen la actividad.

Sección 2: Selección de descriptores y estandarización

Antes de aplicar PCA, se seleccionan únicamente las columnas correspondientes a los descriptores numéricos (MW, LogP, TPSA, HBD, HBA). La variable Activa no se incluye en el cálculo de PCA, porque PCA es un método no supervisado que no utiliza la etiqueta de clase. A continuación, se aplica scale() sobre la matriz de descriptores. Esto genera valores estandarizados (z-scores), donde cada columna queda con media 0 y desviación estándar 1. Esta estandarización es fundamental, ya que impide que una variable con unidades grandes (por ejemplo, MW) domine el análisis frente a otra con rango pequeño (por ejemplo, HBD).

Sección 3: Cálculo de PCA con prcomp()

Con los datos estandarizados, el script utiliza la función prcomp() para calcular el PCA. Esta función obtiene una nueva base de ejes (componentes principales) que son combinaciones lineales de los descriptores originales y que maximizan la varianza explicada: - PC1 es la combinación lineal que explica la mayor parte de la variabilidad total. - PC2 es la segunda combinación lineal más importante, ortogonal (independiente) de PC1. El comando summary(pca) permite ver: - Cuánta varianza explica cada componente. - El porcentaje acumulado de varianza explicada (por ejemplo, PC1 + PC2 juntos). Esto ayuda a decidir si basta con visualizar PC1 y PC2 para tener una buena representación de la estructura de los datos.

Sección 4: Visualización PC1 vs PC2 y actividad biológica

Una vez calculado el PCA, el script extrae las coordenadas de cada molécula en el espacio PC1–PC2 (normalmente usando pca$x[,1] y pca$x[,2]) y genera un gráfico de dispersión donde: - Cada punto corresponde a un compuesto. - El eje X representa PC1. - El eje Y representa PC2. - El color de los

puntos representa la actividad (Activa = Sí/No). Esta visualización permite explorar si los compuestos activos tienden a concentrarse en alguna región particular del espacio de componentes principales, lo que sugeriría que existe una relación entre los descriptores (capturada por PC1 y PC2) y la actividad biológica. Se puede incluir una leyenda, títulos y etiquetas de ejes para facilitar la lectura del gráfico en clases y discusiones grupales.

Sección 5: Interpretación de las cargas (loadings)

El objeto pca$rotation contiene las cargas (loadings) de cada descriptor en cada componente principal. Cada carga indica cuánto contribuye ese descriptor a la dirección del componente. Por ejemplo: - Si PC1 tiene cargas altas (en valor absoluto) para MW y LogP, se puede interpretar que PC1 está relacionado con el “tamaño y lipofilicidad” de los compuestos. - Si PC2 tiene cargas altas para TPSA y HBD, puede interpretarse como un eje de “polaridad y capacidad de formar puentes de hidrógeno”. Analizar estas cargas ayuda a dar significado químico/farmacológico a los componentes principales, en lugar de verlos solo como combinaciones matemáticas abstractas.

Sección 6: Relación entre PC1 y la regla oculta

Una extensión interesante de la práctica consiste en comparar PC1 con el score_lineal utilizado para generar la actividad biológica. Esta comparación se realiza calculando la correlación entre ambos vectores: - Si la correlación es alta y positiva, significa que PC1 está capturando en gran medida la misma combinación de descriptores que define la probabilidad de actividad en la simulación. - Esto ilustra cómo un método no supervisado (PCA), que no ve la etiqueta de actividad, puede recuperar estructura latente que está relacionada con el fenotipo biológico (activo/inactivo). Este análisis refuerza la idea de que PCA no solo reduce dimensiones, sino que también puede revelar patrones latentes que hacen sentido desde el punto de vista químico y farmacéutico.

Sugerencias didácticas

- Pedir al estudiante que describa, en sus propias palabras, qué representa PC1 y qué representa PC2, a partir de las cargas de los descriptores. - Solicitar que interpreten el gráfico PC1–PC2: ¿se separan visualmente los compuestos activos de los inactivos? - Invitar a que comparen esta práctica con los modelos supervisados (Regresión Logística y Random Forest) trabajados previamente, discutiendo similitudes y diferencias entre métodos supervisados y no supervisados. - Como actividad extra, se puede pedir que cambien la “regla oculta” (modificando los coeficientes del score_lineal) para ver cómo afecta la distribución de puntos y la separación visual entre activos e inactivos en el plano PC1–PC2.

Conclusión

Esta práctica integra el uso de PCA con un escenario realista de descriptores moleculares y actividad biológica simulada. El estudiante observa cómo los componentes principales resumen la información de varios descriptores en pocos ejes, facilitan la visualización y pueden relacionarse con propiedades químicas intuitivas y con el comportamiento biológico de los compuestos.