Flujos de trabajo para Resonancia Magnética Nuclear

La Resonancia Magnética Nuclear (RMN) ha permitido grandes avances en metabolómica ya que proporciona una gran cantidad de datos enriquecidos, aunque se requiere un preprocesamiento de señales automatizado antes de aplicar un análisis mediante algoritmos de machine learning o métodos estadísticos. Hay que tener en cuenta que las muestras biológicas pueden contener miles de metabolitos desconocidos por lo que los métodos dirigidos (hablaremos a continuación), basados en un set de metabolitos limitado, no captan la verdadera complejidad de las muestras. Es importante destacar la necesidad de realizar una validación externa siempre que se trabaje en metabolómica computacional para confirmar la validez de los modelos o, en su defecto, utilizar herramientas de doble validación cruzada para dotar al modelo de validez.

Comencemos por el principio. La metabolómica estudia el metaboloma, el conjunto completo de moléculas pequeñas (menos de 1500 Da) que se pueden encontrar en una muestra, y aspira a realizar una caracterización completa y global de las muestras biológicas, pero ello va a depender de las muestras, de su origen y composición variantes. En función de los objetivos del estudio, se puede enfocar la metabolómica desde dos aproximaciones distintas, dirigida y no dirigida.

La metabolómica dirigida parte de un conjunto de biomarcadores preestablecidos y una librería limitada de los espectros de Resonancia Magnética Nuclear (RMN) obtenidos, y se centra en la cuantificación de los metabolitos. La desventaja es que si no se dispone de la librería se pasaría por alto el metabolito. La metabolómica no dirigida, por el contrario, hace una cuantificación de la intensidad de los picos en primer lugar y luego se procede a la identificación a partir de los perfiles de espectros conocidos. Se semicuantifican dichos compuestos y luego se hace un análisis estadístico para encontrar biomarcadores, siendo estos los que pasan a la etapa de identificación de metabolitos. El problema es que puede haber compuestos que se sabe que son discriminantes pero que no se pueden identificar. No todos los metabolitos están en la librería y por tanto nos encontramos señales que no vamos a poder identificar con ninguno de los metabolitos conocidos y caracterizados.

Cuando se aplican técnicas de RMN, los espectrómetros miden la relajación de unos núcleos que vibran a la frecuencia de Larmor después de un pulso de radiofrecuencia (RF). Cuando se acaba el pulso de RF los núcleos siguen oscilando, pero van decayendo. Aplicando la transformada de Fourier (Ernst & Anderson, 1966) obtenemos el espectro de absorción que es lo que nos da la composición química de la muestra (Figura 1). En este caso, la técnica de RMN se aplica a los biofluidos, como la orina, para estudios metabolómicos. Debido a un grado de solapamiento bastante alto, la técnica tiene un rango muy dinámico, complicando la separación de las señales del ruido de fondo. Además, la posición y forma de los picos puede ser distinta por variaciones en el pH y eso dificulta la identificación de los metabolitos de la muestra.

El análisis de datos ha ido creciendo en importancia a lo largo de estos últimos 10 años debido a la progresiva mejora en la instrumentación. Teniendo esto en cuenta, con frecuencia, el cuello de botella es el análisis de estos datos procedentes de RMN y no la obtención de los mismos. Un operador experto necesita entre 30-40 min para analizar un solo espectro, sujeto a error humano y algunas decisiones que tiene que tomar el propio operario, por lo que los datos están sujetos a cierto nivel de subjetividad.

(FIGURA 1) *Figura 1. Fases del procesado de Resonancia Magnética Nuclear. Modificado de Networking Bioinformatics 2020 - Introduction to Computational Metabolomics*

Al completar la fase tres del procesado de RMN se puede dar por finalizada esta etapa, obteniéndose los datos en crudo. Esa información debe ser preprocesada, por lo que se comenzaría con la parte propiamente computacional. Los objetivos de la parte computacional son dos, por un lado el diseño de modelos predictivos y por otro el descubrimiento de marcadores, es decir, no sólo se quieren ver diferencias entre biomarcadores, también se buscan otros nuevos.

Hagamos ahora un recorrido a lo largo de las distintas etapas en el procesamiento computacional de un conjunto de datos, lo que se conoce como flujo de trabajo o workflow. El diseño de un flujo de trabajo dependerá del tipo de datos que se desean procesar, pero en general, todos constan de los siguientes elementos comunes: un dispositivo de medida (instrumentación analítica), un set de ejemplos de calibración para los modelos de entrenamiento y validación de datos, un flujo de trabajo de procesamiento y un algoritmo de clasificación. Teniendo esto en cuenta, para analizar los datos obtenidos a partir de RMN, se sigue un flujo de trabajo concreto cuyo procesamiento consta de varias fases (Figura 2). La primera es la carga de datos del espectro a la cual se le añaden los metadatos. Se realiza una interpolación de los mismos para posteriormente excluir regiones que carecen de utilidad para el estudio. Posteriormente, hay una etapa de detección de puntos aislados mediante análisis de componentes principales robustos y la corrección de la línea de base.

Se procede a una detección de los picos (uno de los algoritmos que pueden usarse es el algoritmo CluPA, basado en un agrupamiento en una dimensión de las posiciones de los picos utilizando como distancia la correlación, calculada de forma rápida en el dominio de Fourier). El espectro como tal no puede ser usado debido a la cantidad de puntos que se pueden encontrar y a la gran dimensionalidad, lo cual conllevaría una gran cantidad de problemas a la hora de aplicar los métodos de machine learning. Lo ideal es utilizar binning, que consiste en dividir el espectro en segmentos yuxtapuestos para disminuir las dimensiones. La división de los mismos se realiza de forma que se adapten a los picos para evitar inestabilidades en las áreas de cada uno de los segmentos.

Para poder conseguirlo, es necesario una detección correcta de los picos de señal y su integración mediante el uso del algoritmo Wavelet, el cual posee gran robustez contra ruidos, distorsiones en la forma de los picos y problemas de la línea de bases. La integración es esencial, ya que errores en este paso podrían dar lugar a una interpretación errónea de los resultados en fases posteriores del análisis. Una vez identificados los picos se procede a una normalización digital o numérica, su integración y la obtención de la tabla de datos lista para aplicar los métodos de machine learning.

(FIGURA 2) *Figura 2. Workflow de preprocesamiento: (1) Carga de datos del espectro, (2) Adición de los metadatos, (3) Interpolación, (4) Exclusión de regiones no interesantes, (5) Etapa de detección de outliers mediante análisis de componentes principales robustos (y corrección de la línea de base), (6) Detección de los picos por muestra, (7) Alineamiento de picos (es un punto crítico para asegurar que se realiza un peak matching), (8) Normalización, (9) Integración de picos, (10) Obtención de la tabla de datos para Machine Learning. Imagen obtenida de Networking Bioinformatics 2020 - Introduction to Computational Metabolomics.*

Después del preprocesamiento de datos del espectro, se obtiene una tabla que contiene los datos en crudo. Sin embargo, dichos datos todavía no están listos para su análisis mediante modelos de machine learning. Para ello, se requiere realizar una serie de pasos previos a la aplicación de los algoritmos clasificadores. El preprocesado de los datos en crudo suele incluir un paso de autoescalado, necesario para compensar la diferencia de intensidad entre los picos. Sin éste, se pasarían por alto los picos de intensidad más baja cercanos a otros picos más altos. Además, el preprocesado suele incluir una transformación no lineal, normalmente logarítmica o exponencial, para intentar mejorar la distribución gaussiana de las intensidades. A continuación se procede al análisis exploratorio de componentes principales y a la eliminación de puntos aislados para finalmente aplicar los clasificadores.

A continuación, se pueden aplicar diversos modelos de predicción. Los algoritmos que más se utilizan para análisis de datos de RMN son tres, por un lado el análisis discriminante por mínimos cuadrados parciales (PLS-DA), un método lineal basado en una reducción de la dimensionalidad orientada. Este algoritmo tiene una alta tendencia al sobreajuste, incluso cuando se realiza una validación cruzada. Por ello, en general son optimistas en exceso, así que siempre se requiere validación externa para PLS-DA. Por otra parte, se utiliza también el conocido como Random Forest, un método no lineal que además ofrece una clasificación de variables. Y por último, se puede aplicar el SVM (Support Vector Machine), que presenta como ventaja una baja tendencia al sobreajuste. Es muy importante la validación externa de los modelos (Marco, 2014), siendo recomendable utilizar muestras ciegas para ello, fuera de ese set de datos. Aunque en el caso de disponer de pocas muestras, existe la opción de realizar una doble validación cruzada que nos permitiría tomar como ciegas a las propias muestras del ensayo.

En metabolómica, existe una tendencia general al uso de herramientas estadísticas de test de hipótesis para descubrir biomarcadores. Sin embargo, mediante este método no se tienen en cuenta la interacción entre características y se requiere una corrección multitest a posteriori, por lo que se acaba perdiendo sensibilidad.

La alternativa es el descubrimiento de biomarcadores multivariados, que se pueden dividir en dos grandes familias (figura 3). Por un lado, los filtros son métodos iterativos que poseen un algoritmo de búsqueda automática basado en una función objetivo, capaz de especificar a priori el potencial predictivo de cada característica antes de añadirlo a un algoritmo de machine learning confiriendo rapidez a esta técnica. Por ejemplo, la aplicación de la puntuación de Fisher indica si esa característica será informativa.

Por otro lado, los wrappers son métodos que iteran sobre el propio algoritmo de machine learning en búsqueda de un subconjunto de características con mejor poder predictivo. Si el modelo de predicción es muy costoso a nivel de entrenamiento, es decir, necesita una cantidad de tiempo considerable para entrenarse, el proceso se ralentiza notablemente debido a las numerosas iteraciones que se requieren.

También son muy populares los algoritmos híbridos los cuales aportan una clasificación de variables. Trabajan de forma similar a los filtros, pero su función objetivo está basada en una lista proporcionada por el propio algoritmo de machine learning. Esto confiere rapidez al método ya que se hace en una única iteración. Tanto para la parte de preprocesado como para la parte de machine Learning, una herramienta útil es el paquete AlpsNMR, que ofrece la ventaja de proveer de un p-valor a cada uno de los VIP (Variable Importance in Projection) mediante el uso de test de permutaciones. Como alternativa a AlpsNMR, se puede utilizar el paquete MUVR para R en el cual viene implementado un método que realiza la clasificación de variables dentro

del mismo algoritmo de búsqueda. En este caso, se genera una clasificación de variables y se elimina el 20% menos significativo. A continuación se repite el proceso de búsqueda, se aplica el algoritmo clasificador y se genera una nueva clasificación de la que se descartaría de nuevo un 20%. El bucle se repite hasta que se comienza a perder poder predictivo.

Como hemos recogido en este artículo, la implementación de flujos de trabajo complejos nos permite realizar identificaciones del metaboloma presente en muestras biológicas muy diversas, todo un reto al que la metabolómica se enfrenta con ayuda de la RMN. Este artículo ha pretendido transmitir los principales puntos expuestos en la conferencia educativa " Introduction to Computational Metabolomics", impartida por el Dr. Santiago Marco en el último Networking – Bioinformatics organizado por My Scientific.

BIBLIOGRAFÍA

Ernst, R. R., Anderson, W. A. Application of Fourier transform spectroscopy to magnetic resonance. Review of Scientific Instruments, 37, 93-102 (1966). https://doi.org/10.1063/1.1719961

Marco, S. The need for external validation in machine olfaction: emphasis on health-related applications. Analytical and Bioanalytical Chemistry, 406, 3941–3956 (2014). https://doi.org/10.1007/s00216-014-7807-7