UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Statistical methods development for the multiomic systems biology

Author: Manuel Ugidos Guerrero

Supervisors: Dr. Sonia Tarazona Campos

Dr. Ana Conesa Cegarra

Dr. Alberto J. Ferrer Riquelme

Resumen

La investigación en Biología de Sistemas se ha expandido en los últimos años junto con el desarrollo de las tecnologías ómicas. La combinación y el análisis simultáneo de diferentes tipos de datos ómicos permite el estudio de las conexiones y relaciones entre los diferentes niveles de organización celular, lo cuál permite una visión a nivel de sistema del organismo estudiado. La presente tesis doctoral tiene como objetivo estudiar, desarrollar y aplicar estrategias de integración multiómica al campo de la biología de sistemas.

El todavía elevado coste de las tecnologías ómicas, dificulta que la mayoría de laboratorios puedan abordar un estudio multiómico completo. No obstante, la gran disponibilidad de datos ómicos en repositorios públicos, permite el uso de estos datos ya generados. Desafortunadamente, la combinación de datos ómicos provenientes de diferentes orígenes, da lugar a la aparición de un ruido no deseado en los datos, lo que se conoce como efecto lote o *"batch effect"* en inglés. El efecto lote impide el correcto análisis conjunto de los datos y, por lo tanto, es necesario el uso de los llamados Algoritmos de Corrección de Efecto Lote para eliminarlo. En la actualidad, existe un gran número de éstos algoritmos que corrigen el efecto lote que se basan en diferentes métodos y modelos estadísticos, y que forman parte del paso de preprocesado de los datos. Sin embargo, los métodos existentes no están pensados para los diseños multiómicos ya que solo permiten la cor-

rección de un mismo tipo de dato ómico que debe haber sido medido en todos los lotes o batches. Por esta razón desarrollamos nuestra herramienta MultiBaC basada en la regresión PLS y modelos ANOVA-SCA, que permite la corrección del efecto lote en diseños multiómicos, permitiendo la corrección de datos que no hayan sido medidos en todos los lotes. En este trabajo, MultiBaC fué validado y evaluado en diferentes conjuntos de datos, además presentamos MultiBaC como paquete de R para facilitar el uso de nuestra herramienta.

La mayoría de métodos existentes de integración multiómica son métodos multivariantes basados en el análisis del espacio latente. Estos métodos se conocen como "dirigidos por datos" o *"data-driven"* en inglés. Este tipo de métodos se basan en la búsqueda de correlaciones para determinar las relaciones entre las distintas variables. Los métodos dirigidos por datos necesitan de gran cantidad de observaciones o muestras para poder encontrar correlaciones robustas y/o significativas entre las variables. Lamentablemente, en el mundo de la biología molecular, los conjuntos de datos con un gran número de muestras no suelen ser muy habituales, debido de nuevo al elevado coste de generación de los datos ómicos. Como alternativa a los métodos dirigidos por datos, algunas estrategias de integración multiómicas se basan en métodos "dirigidos por modelos" o *"model-driven"* en inglés. Estos métodos pueden ajustarse con un menor número de observaciones y son muy útiles para encontrar relaciones mecanísticas entre los diferentes componentes celulares. Sin embargo, los métodos dirigidos por mode-

los necesitan de una información a priori, el modelo, que normalmente es un modelo metabólico del organismo estudiado. Actualmente, únicamente transcriptómica y metabolómica cuantitativa, han sido los dos tipos de dato ómico que se han integrado con éxito usando métodos dirigidos por modelos. No obstante, la metabolómica cuantitativa no está muy extendida y la mayoría de laboratorios generan metabolómica no cuantitativa o semi-cuantitativa, las cuáles no pueden integrarse con los métodos actuales. Para contribuir en esta cuestión, desarrollamos MAMBA, una herramienta de integración multiómica dirigida por modelos y basada en métodología de optimización matemática, que es capaz de analizar conjuntamente metabolómica no cuantitativa o semi-cuantitativa con otro tipo de ómica asociada a genes, como por ejemplo la trascriptómica. MAMBA fue comparado con otros métodos existentes en cuanto a la capacidad de predcción de metabolitos y fué aplicado al conjunto interno de datos multiómicos. Este conjunto de datos multiómicos fue generado dentro del proyecto PROMETEO, en el cuál está enmarcada esta tesis. MAMBA demostró capturar la biología conocida sobre nuestro diseño experimental, además de ser útil para derivar nuevas observaciones e hipótesis biológicas.

En conjunto, esta tesis presenta herramientas útiles para el campo de la biología de sistemas, y que cubren tanto el preprocesado de conjunto de datos multiómicos como su posterior análisis estadístico integrativo.