Las variables ómicas en abundancias relativas son (casi casi) coherentes en la subcomposición

Marina Martínez-Álvaro[1], Michael Greenacre[2], Agustín Blasco[1]*

[1]Instituto de Ciencia y Tecnología Animal, Camino de Vera s/n Edificio 7G, 46020, Valencia
[2]Departamento de Economía y Empresa, Universitat Pompeu Fabra, Carrer de Ramon Trias Fargas, 25-27, 08005, Barcelona
* Corresponding author: mamaral9@upv.es

Resumen

Muchas bases de datos ómicas (microbioma, transcriptoma, metaboloma) son composicionales debido a que múltiples pasos en su medición implica *subsettings*, cada uno generando una nueva restricción muestral: la toma de muestras, la extracción de DNA, la amplificación o la secuenciación. Los counts obtenidos contienen solo información relativa (la escala total se pierde por completo), y la suma total de counts generada por muestra no está relacionada con las cantidades absolutas en la matriz original, sino que más bien tiene que ver con aspectos técnicos de la eficiencia en la extracción o la sensibilidad del secuenciador. La eliminación de esta variación de naturaleza sistemática se puede hacer por varios métodos; uno de los más sencillos es la normalización TSS, donde cada muestra se divide por su suma total de counts y se expresa, por tanto, en abundancias relativas. Uno de los principales problemas de los datos relativos es la falta de coherencia subcomposicional, es decir, que las abundancias relativas cambian cuando los datos se renormalizan tras eliminar o añadir variables. Aunque este problema está bien documentado en composiciones pequeñas, no se ha investigado en bases de datos ómicas, que normalmente contienen cientos o miles de variables, y en las que las subcomposiciones ocurren constantemente. Las subcomposiciones surgen, por ejemplo, al resecuenciar con distinta profundidad, al utilizar distintas bases de datos de referencia para pasar de reads a counts, o al filtrar la base de datos final por variables con muchos ceros o por variables que no se han podido identificar. La falta de coherencia subcomposicional afecta gravemente a la reproducibilidad de los resultados. La solución estándar a este problema es el uso de transformaciones log-ratio entre cada par de variables, cuyos valores se mantienen, obviamente, en cualquier subcomposición, pero en el caso de las ómicas esta transformación genera una cantidad de variables log-ratio imposible de interpretar, y además requiere tratar los ceros, que pueden llegar a ser hasta del 80–90 % en bases de datos ómicas muy dispersas, y cuya imputación introduce variabilidad espuria.

En este trabajo hemos evaluado la coherencia subcomposicional en cinco bases de datos ómicas habituales en nuestro campo: secuenciación 16S de heces, secuenciación metagenómica completa ruminal (a nivel taxonómico y funcional), transcriptómica hepática y metabolómica plasmática, con transformación TSS. Hemos hecho una evaluación exhaustiva considerando tanto contextos de aprendizaje no supervisado como supervisado, lineal y no lineal. Brevemente, generamos 100 subcomposiciones aleatorias que comprendían un tercio de las variables originales y comparamos sus resultados estadísticos con los de la composición completa. Los datos ómicos relativos mostraron una coherencia casi perfecta: las abundancias relativas, las correlaciones, covariables y distancias entre variables, y las distancias entre muestras, presentaron concordancias muy altas (≥ 0,98–0,99). Los resultados de modelos supervisados (análisis diferencial, regresión lineal, PLS, random forest y modelos lineales mixtos) también fueron altamente coherentes subcomposicionalmente. Concluimos que las bases de datos ómicas en valores relativos son suficientemente coherentes para fines prácticos, refutando así una de las principales críticas al uso de abundancias relativas en el campo de la ómica frente a las transformaciones log-ratio

Keywords: datos compositionales, ómicas, coherencia subcompositional, aprendizaje supervisado; aprendizaje no supervisado; transformaciones log-ratio

Agradecimientos: Agradecemos a Scotland's Rural College (SRUC, Edimburgo, Reino Unido) por compartir datos metagenómicos financiados por el Gobierno de Escocia; BBSRC (BB/N01720X/1, BB/N016742/1, BB/S006567/1 y BB/S006680/1); Genus plc; AHDB; y QMS. Marina Martínez-Álvaro

agradece al Ministerio de Ciencia e Innovación de España por una ayuda Ramón y Cajal (RYC2021-032618-I) financiada por MCIN/AEI/10.13039/501100011033 y por la Unión Europea NextGenerationEU/PRTR. Los autores agradecen el apoyo financiero del programa APICID de la Universitat Politècnica de València para estancias docentes de corta duración fuera de Europa y la invitación de profesores extranjeros para realizar actividades docentes breves.